

CORRETTEZZA E MACHINE LEARNING

Limitazioni e opportunità

Solon Barocas, Moritz Hardt, Arvind Narayanan

Contenuto

Prefazione	v
Ringraziamenti	x
1 Introduzione	1
Disparità demografiche	3
Il ciclo di apprendimento automatico.	4
Lo stato della società.	5
Il problema della misurazione.	7
Dai dati ai modelli.	10
Le insidie dell'azione.	12
Feedback e cicli di feedback.	13
Diventare concreto con un esempio di giocattolo.	15
Giustizia oltre il giusto processo decisionale.	18
La nostra prospettiva: limiti e opportunità.	20
Note bibliografiche e approfondimenti.	21
2 Quando il processo decisionale automatizzato è legittimo?	23
L'apprendimento automatico non sostituisce il processo decisionale umano.	24
La burocrazia come baluardo contro il processo decisionale arbitrario.	25
Tre forme di automazione.	27
Discrepanza tra target e obiettivo.	34
Non prendere in considerazione le informazioni rilevanti.	35
I limiti dell'induzione.	38
Il diritto a previsioni accurate?	39
Agenzia, ricorso e colpevolezza.	40
Considerazioni conclusive.	43
3 Classificazione	44
Modellazione delle popolazioni come distribuzioni di probabilità.	44
Classificazione formalizzante.	46
Apprendimento supervisionato.	50
Gruppi nella popolazione.	52
Criteri statistici di non discriminazione	54
Indipendenza.	55

Separazione.	56
Sufficienza.	60
Come soddisfare un criterio di non discriminazione.	63
Relazioni tra criteri.	64
Caso di studio: punteggio del credito.	67
Limitazioni intrinseche dei criteri di osservazione.	71
Note del capitolo.	73
4 Nozioni relative di equità	76
Svantaggio relativo sistematico.	76
Sei resoconti dell'illegittimità della discriminazione.	78
Intenzionalità e discriminazione indiretta.	80
Pari opportunità.	81
Tensioni tra le diverse visioni.	85
Merito e deserto.	88
Il costo dell'equità.	90
Collegamento tra nozioni statistiche e morali di equità.	92
I fondamenti normativi della parità del tasso di errore.	97
Alternative per realizzare la visione mediana delle pari opportunità.	101
Riepilogo	101
5 Causalità I	104
limiti dell'osservazione	105
Modelli causali.	107
Grafici causali.	111
Interventi ed effetti causali.	113
Confondente.	114
Analisi grafica della discriminazione.	117
Controfattuali.	121
Analisi controfattuale della discriminazione.	127
Validità del modello causale.	132
Note del capitolo.	137
6 Comprendere la legge antidiscriminatoria degli Stati Uniti	139
Storia e panoramica della legge antidiscriminatoria statunitense.	140
Alcune nozioni di base del sistema giuridico americano.	146
Come la legge concepisce la discriminazione.	151
Limiti della legge nel contrasto alle discriminazioni.	155
Regolamentare l'apprendimento automatico.	160
Considerazioni conclusive.	169
7 Testare la discriminazione nella pratica	171
Parte 1: Test tradizionali di discriminazione	172
Studi di audit.	172
Testare l'impatto dell'accecamento.	175
Rivelare fattori estranei alle decisioni.	177

Testare l'impatto delle decisioni e degli interventi.	178
Test puramente osservativi.	179
Riepilogo dei test e dei metodi tradizionali.	182
Discriminazione basata sul gusto e statistica.	184
Studi sui processi decisionali e sulle organizzazioni.	186
Parte 2: Testare la discriminazione nei sistemi algoritmici	187
Considerazioni sull'equità nelle applicazioni dell'elaborazione del linguaggio naturale.	188
Disparità demografiche e applicazioni discutibili della visione artificiale	191
Sistemi di ricerca e raccomandazione: tre tipologie di dati.	192
Comprendere l'ingiustizia nel targeting degli annunci.	194
Considerazioni sull'equità nella progettazione dei mercati online.	196
Meccanismi di discriminazione.	198
Criteri di equità negli audit algoritmici.	199
Flusso delle informazioni, correttezza, privacy.	201
Confronto dei metodi di ricerca.	202
Guardando avanti.	204
Note del capitolo.	204
8 Una visione più ampia della discriminazione	205
Caso di studio: il divario salariale di genere su Uber	205
Tre livelli di discriminazione.	209
Apprendimento automatico e discriminazione strutturale.	213
Interventi strutturali per il machine learning equo.	217
Interventi organizzativi per un processo decisionale più giusto.	221
Note del capitolo.	229
Appendice: uno sguardo più approfondito ai fattori strutturali.	230
9 Set di dati	232
Una panoramica dei set di dati in diversi domini	233
Ruoli giocati dai set di dati.	240
Danni associati ai dati.	250
Oltre i set di dati.	254
Riepilogo	261
Note del capitolo.	261

Prefazione

Un modo peculiare di prendere decisioni è caratteristico della società moderna. Le istituzioni di tutti i tipi, dalle aziende ai governi, rappresentano le popolazioni come tabelle di dati. Le righe fanno riferimento a individui. Le colonne contengono misurazioni su di esse. I macchinari statistici applicati a queste tabelle consentono ai loro proprietari di estrarre modelli che si adattano all'aggregato.

Poi arriva un atto di fede. Dobbiamo immaginare che risultati sconosciuti, futuri o non osservati, nel percorso di vita di un individuo seguano gli schemi che ha trovato. Dobbiamo accettare le decisioni prese come se tutti gli individui seguissero la regola dell'aggregato. Dobbiamo fingere con noi stessi che guardare al futuro sia guardare al passato. È un atto di fede che da secoli è alla base di decisioni consequenziali. Alimentato dai primi successi nella determinazione dei prezzi assicurativi e nella valutazione del rischio finanziario, il processo decisionale statistico di questo tipo si è fatto strada in quasi tutti gli aspetti della nostra vita. Ciò che ne ha accelerato l'adozione negli ultimi anni è stata la crescita esplosiva dell'apprendimento automatico, spesso sotto il nome di intelligenza artificiale.

L'apprendimento automatico condivide fondamenti teorici delle decisioni consolidati da tempo con gran parte della statistica, dell'economia e dell'informatica. Ciò che aggiunge l'apprendimento automatico è un repertorio di euristiche in rapida crescita che trovano regole decisionali da set di dati sufficientemente grandi. Queste tecniche per adattare enormi modelli statistici su grandi set di dati hanno portato a diversi risultati tecnologici impressionanti. La classificazione delle immagini, il riconoscimento vocale e l'elaborazione del linguaggio naturale hanno fatto passi da gigante. Sebbene questi progressi spesso non siano direttamente correlati a specifici contesti decisionali, danno forma a narrazioni sulle nuove capacità dell'apprendimento automatico.

Per quanto utile sia l'apprendimento automatico per alcune applicazioni positive, viene utilizzato con grande efficacia anche per il tracciamento, la sorveglianza e la guerra. Dal punto di vista commerciale, i suoi casi d'uso di maggior successo fino ad oggi sono la pubblicità mirata e la raccomandazione di contenuti digitali, entrambi di discutibile valore per la società. Dalle sue radici nella cibernetica e nella teoria del controllo dell'era della Seconda Guerra Mondiale, l'apprendimento automatico è sempre stato politico. I progressi nell'intelligenza artificiale alimentano un complesso militare industriale globale e sono finanziati da esso. Le storie di successo raccontate sull'apprendimento automatico supportano anche coloro che vorrebbero adottare algoritmi in domini esterni a quelli studiati dagli informatici. Un mercato opaco di fornitori di software fornisce strumenti decisionali algoritmici da utilizzare nelle forze dell'ordine, nella giustizia penale, nell'istruzione e nei servizi sociali. In molti casi ciò che viene commercializzato e venduto

poiché l'intelligenza artificiale sono metodi statistici che praticamente non sono cambiati da decenni.

Molti danno per scontato il atto di fede dietro il processo decisionale statistico a un punto tale che è diventato difficile metterlo in discussione. Intere discipline hanno abbracciato modelli matematici del processo decisionale ottimale nei loro fondamenti teorici.

Gran parte della teoria economica considera le decisioni ottimali come un presupposto e un ideale del comportamento umano. A loro volta, altre discipline etichettano le deviazioni dall'ottimalità matematica come "bias" che invita all'eliminazione. Volumi di articoli accademici parlano degli evidenti pregiudizi dei decisori umani.

In questo libro, prendiamo l'apprendimento automatico come motivo per rivisitare questo atto di fede e per interrogarci su come le istituzioni prendono decisioni sugli individui. Il processo decisionale istituzionale è stato a lungo formalizzato tramite procedure burocratiche e l'apprendimento automatico ha molto in comune con esso. In molti casi, l'apprendimento automatico viene adottato per migliorare e talvolta automatizzare le decisioni ad alto rischio prese abitualmente dalle istituzioni. Pertanto, non confrontiamo i modelli di apprendimento automatico con i giudizi soggettivi dei singoli esseri umani, ma piuttosto con il processo decisionale istituzionale . Interrogare l'apprendimento automatico è un modo per interrogare il processo decisionale istituzionale nella società di oggi e nel prossimo futuro.

Se l'apprendimento automatico è il nostro modo di studiare il processo decisionale istituzionale, l'equità è la lente morale attraverso la quale esaminiamo tali decisioni. Gran parte della nostra discussione si applica a scenari concreti di screening, selezione e allocazione.

Un tipico esempio è quello di un datore di lavoro che accetta o rifiuta candidati per un posto di lavoro. Un modo per interpretare l'equità in tali scenari decisionali è come l'assenza di discriminazione. Questa prospettiva è micro nella misura in cui gli individui sono l'unità di analisi. Studiamo come le caratteristiche misurate di un individuo portano a risultati diversi. Gli individui sono l'elemento costitutivo sociologico. Una popolazione è un insieme di individui. I gruppi sono sottoinsiemi della popolazione. Un decisore ha il potere di accettare o rifiutare gli individui per l'opportunità che cercano.

La discriminazione in questa visione riguarda un'errata considerazione sulla base dell'appartenenza al gruppo . Il problema riguarda tanto cosa significa illecito quanto cosa è alla base di. Anche la discriminazione non è un concetto generale. È un settore specifico in quanto si riferisce alle opportunità che influenzano la vita delle persone. Si occupa di categorie socialmente rilevanti che sono servite da base per trattamenti ingiustificati e sistematicamente avversi.

Il primo capitolo dopo l'introduzione esplora le proprietà che rendono il processo decisionale automatizzato una questione di interesse normativo significativo e unico.
In particolare, collichiamo la nostra esplorazione dell'apprendimento automatico in una storia più lunga di riflessione critica sui pericoli del processo decisionale burocratico e sulla sua applicazione meccanica di regole formalizzate. Prima ancora di dedicarci alla questione della discriminazione, chiediamo innanzitutto cosa rende legittimo il processo decisionale automatizzato. Così facendo, isoliamo le proprietà specifiche dell'apprendimento automatico che lo distinguono da altre forme di automazione lungo una serie di dimensioni normative.

A partire dagli anni '50, gli studiosi hanno sviluppato modelli formali di discriminazione che descrivono il trattamento ineguale di più gruppi diversi nella popolazione da parte di un decisore. Nel capitolo 3 ci addentreremo nella teoria delle decisioni statistiche, permettendo

di formalizzare una serie di criteri di equità. I criteri di equità statistica esprimono diverse nozioni di uguaglianza tra gruppi. Riduciamo il vasto spazio delle definizioni formali essenzialmente a tre diverse definizioni reciprocamente esclusive. Ogni definizione risuona con una diversa intuizione morale. Nessuna è sufficiente a supportare affermazioni conclusive di equità. Né queste definizioni sono obiettivi adatti per l'ottimizzazione. Soddisfare uno di questi criteri consente soluzioni palesemente ingiuste. Nonostante i loro limiti significativi, queste definizioni hanno avuto influenza nel dibattito sull'equità.

Il capitolo 4 esplora i fondamenti normativi delle obiezioni alle differenze sistematiche nel trattamento di diversi gruppi e alle disuguaglianze nei risultati sperimentati da questi gruppi. Esaminiamo i numerosi resoconti sull'illegittimità della discriminazione e mostriamo come questi si collegano a diversi punti di vista su cosa significherebbe garantire pari opportunità. In tal modo, evidenziamo alcune tensioni tra visioni contrastanti di pari opportunità – alcune piuttosto ristrette e altre piuttosto ampie – e i vari argomenti che sono stati avanzati per aiutare a risolvere questi conflitti. Detto questo, esploreremo poi come le intuizioni morali comuni e le teorie morali consolidate possano aiutarci a dare un senso ai formalismi introdotti nel capitolo 3, con l'obiettivo di dare a queste definizioni maggiore sostanza normativa.

Negli studi sia tecnici che giuridici sulla discriminazione è presente l'idea di assegnare un peso normativo alle relazioni causali. L'appartenenza al gruppo è stata la causa del rifiuto? Il richiedente sarebbe stato respinto se fosse stato di una razza diversa? Sarebbe stata accettata se non fosse stato per il suo genere? Per comprendere questo tipo di affermazioni e il ruolo che la causalità gioca nella discriminazione, il capitolo 5 di questo libro è un'introduzione autonoma ai concetti formali di causalità.

Dopo il nostro incontro formale con le definizioni di equità, sia statistiche che causali, ci rivolgiamo alle dimensioni legali della discriminazione negli Stati Uniti nel capitolo 6. La situazione legale non si adatta in modo chiaro ai fondamenti morali né al lavoro formale, complicando considerevolmente la situazione. Le due dottrine giuridiche dominanti, trattamento disparato e impatto disparato, sembrano creare una tensione tra la considerazione esplicita dell'appartenenza al gruppo e l'intervento per evitare la discriminazione.

Estendendosi sia al capitolo causale che a quello legale, il capitolo 7 approfondisce in dettaglio le complessità dei test di discriminazione nella pratica attraverso esperimenti e audit.

Lo studio della discriminazione nel processo decisionale è stato criticato come una prospettiva ristretta su un sistema più ampio di ingiustizia per almeno due ragioni. In primo luogo, come nozione di discriminazione trascura i potenti determinanti strutturali della discriminazione, come le leggi e le politiche, le infrastrutture e l'istruzione. In secondo luogo, orienta lo spazio di intervento verso soluzioni che riformano i sistemi decisionali esistenti, nel caso dell'apprendimento automatico tipicamente tramite aggiornamenti di un algoritmo. In quanto tale, la prospettiva può sembrare dare priorità alle "correzioni tecnologiche" rispetto a interventi strutturali più potenti e alternative all'implementazione complessiva di un sistema di apprendimento automatico.

Piuttosto che prevedere la mancata comparizione in tribunale e punire gli imputati per questo, ad esempio, forse l'intervento migliore è facilitare l'accesso alle audizioni giudiziarie.

punti fornendo trasporto e assistenza all'infanzia. Il capitolo 8 introduce il lettore a questa prospettiva più ampia e allo spazio di interventi ad essa associato da un punto di vista empirico.

Riconoscendo l'importanza di una prospettiva sociale e strutturale più ampia, perché dovremmo continuare a studiare la nozione di discriminazione nel processo decisionale? Un vantaggio è che fornisce una strategia politica e legale per esercitare pressioni sui singoli decisori. Possiamo avanzare denunce di discriminazione contro una specifica persona, azienda o istituzione. Possiamo discutere quali interventi esistano in una ragionevole vicinanza al decisore e che quindi ci aspettiamo che il decisore implementi. Alcuni di questi microinterventi potrebbero anche essere più direttamente realizzabili rispetto agli interventi strutturali.

Assumere una prospettiva micro non significa sicuramente ignorare il contesto. In effetti, le regole di allocazione che evitano la considerazione esplicita dell'appartenenza al gruppo creando allo stesso tempo opportunità per un gruppo probabilmente lo fanno collegando la regola di allocazione con fatti sociali esterni. Un esempio importante è la "regola del dieci per cento del Texas" che garantisce agli studenti del Texas che si sono diplomati nel dieci per cento più alto della loro classe di scuola superiore l'ammissione automatica a tutte le università finanziate dallo stato. La norma non sarebbe efficace nel promuovere la diversità razziale nei campus universitari pubblici se le classi delle scuole superiori non fossero segregate fin dall'inizio. Questo esempio illustra che non esiste una mutua esclusività tra l'esame dettagliato di specifiche regole decisionali e il prestare attenzione al contesto sociale più ampio. Piuttosto questi vanno di pari passo.

Un consequenziale punto di contatto tra il mondo sociale più ampio e l'ecosistema del machine learning sono i set di dati. Un capitolo intero esplora la storia, il significato e le basi scientifiche dei set di dati di machine learning. Una considerazione dettagliata dei set di dati, della raccolta e della costruzione dei dati, nonché dei dati associati ai dati tendono a mancare nei programmi di apprendimento automatico.

L'equità rimane un'area di ricerca attiva e tutt'altro che consolidata. Abbiamo scritto questo libro in un periodo di attività di ricerca esplosiva. Migliaia di articoli correlati sono apparsi negli ultimi cinque anni di scrittura. Molti di loro propongono interventi algoritmici che promuovono l'equità. Questo testo non è un'indagine su quest'area in rapida evoluzione, né è un riferimento definitivo. Il capitolo finale, disponibile online, fornisce un punto di ingresso alla ricerca emergente sugli interventi algoritmici.

Il libro presenta alcuni limiti seri, forse evidenti.

Gran parte del nostro libro riguarda specificamente gli Stati Uniti. Scritto da tre autori formati e impiegati presso istituzioni statunitensi, il libro si basa sulla tradizione morale occidentale, presuppone le leggi e la teoria giuridica degli Stati Uniti e fa riferimento al contesto industriale e politico degli Stati Uniti in ogni sua parte. Non abbiamo fatto alcun tentativo di affrontare questa grave limitazione in questo libro. In effetti, ci vorrebbe un libro completamente diverso per affrontare questa limitazione.

Una seconda limitazione deriva dal fatto che il nostro obiettivo primario era sviluppare le basi morali, normative e tecniche necessarie per affrontare l'argomento. Grazie alla sua attenzione ai fondamenti, il libro sembrerà ad alcuni un passo lontano dalle importanti esperienze di quegli individui e comunità più gravemente danneggiati e danneggiati dall'uso degli algoritmi. Questa lacuna è aggravata dal fatto che gli autori di questo libro non hanno esperienza diretta dei sistemi di

oppressione di cui gli algoritmi fanno parte. Di conseguenza, questo libro non sostituisce il lavoro vitale di quegli attivisti, giornalisti e studiosi che ci hanno insegnato i pericoli del processo decisionale algoritmico nel contesto. Ci basiamo su questi contributi essenziali per scrivere questo libro. Abbiamo mirato a evidenziarli ovunque, anticipando che probabilmente non saremmo stati all'altezza in alcuni modi significativi.

Il libro non è né un sostegno totale al processo decisionale algoritmico, né un ampio atto d'accusa. Nello scrivere questo libro, tentiamo quella che è probabilmente la posizione meno popolare su qualsiasi argomento: un equilibrio. Cerchiamo di capire dove il processo decisionale algoritmico ha merito, dedicando al contempo un'attenzione significativa ai suoi danni e limiti. Alcuni vedranno il nostro equilibrio come una mancanza di impegni politici, una sorta di bilateralismo.

Nonostante l'urgenza della situazione politica, il nostro libro non fornisce una guida pratica diretta per prendere decisioni giuste. È un dato di fatto, abbiamo scritto questo libro per un lungo periodo. Siamo convinti che i dibattiti sul processo decisionale algoritmico persisteranno. Il nostro obiettivo è rafforzare le basi intellettuali dei dibattiti futuri, che si svolgeranno in migliaia di casi specifici. Chiunque spera di dare forma a questo futuro del processo decisionale algoritmico nella società probabilmente troverà materiale utile in questo libro.

Alcuni capitoli, in particolare il capitolo 3 sulla classificazione e il capitolo 5 sulla causalità, richiedono prerequisiti matematici significativi, principalmente in probabilità e statistica per gli studenti universitari. Tuttavia, gli altri capitoli li dedichiamo a un pubblico molto più ampio. Ci auguriamo che gli studenti di più campi trovino questo libro utile nella preparazione alla ricerca in aree correlate. Il libro non rientra perfettamente nei confini disciplinari di nessun singolo dipartimento. Di conseguenza offre ai lettori l'opportunità di andare oltre i programmi di studio stabiliti nella loro disciplina primaria.

Da quando abbiamo iniziato a pubblicare il materiale tratto da questo libro, anni fa, gli istruttori hanno incorporato il materiale in una varietà di corsi, sia a livello universitario che universitario, in diversi dipartimenti. Centinaia di lettori ci hanno inviato feedback estremamente utili di cui siamo profondamente grati.

E a coloro che si lamentano dei nostri lenti progressi nello scrivere questo libro, rispondiamo con empatia: è giusto.

Ringraziamenti

Questo libro non sarebbe stato possibile senza il profondo contributo dei nostri collaboratori e della comunità in generale.

Siamo grati ai nostri studenti per la loro partecipazione attiva ai corsi pilota a Berkeley, Cornell e Princeton. Grazie in particolare a Claudia Roberts per gli appunti delle lezioni del corso di Princeton.

Un ringraziamento speciale a Katherine Yen per l'aiuto editoriale e tecnico con il libro.

Moritz Hardt è debitore a Cynthia Dwork per averlo introdotto al tema di questo libro durante uno stage formativo nel 2010.

Abbiamo beneficiato di discussioni, feedback e commenti sostanziali da parte di Rediet Abebe, Andrew Brunskill, Aylin Caliskan, André Cruz, Frances Ding, Michaela Hardt, Lily Hu, Ben Hutchinson, Shan Jiang, Sayash Kapoor, Lauren Kaplan, Niki Kilbertus, Been Kim, Kathy Kleiman, Issa Kohler-Hausmann, Mihir Kshirsagar, Eric Lawrence, Zachary Lipton, Lydia T. Liu, John Miller, Smitha Milli, Shira Mitchell, Jared Moore, Robert Netzorg, David Parkes, Juan Carlos Perdomo, Eike Willi Petersen, Daniele Regoli, Ofir Reich, Claudia Roberts, Olga Russakovsky, Matthew J. Salganik, Carsten Schwemmer, Ludwig Schmidt, Andrew Selbst, Matthew Sun, Angelina Wang, Christo Wilson, Annette Zimmermann, Tijana Zrnic.

Arvind Narayanan è grato per il sostegno della National Science Foundation nell'ambito delle sovvenzioni IIS-1763642 e CHS-1704444.

1

Introduzione

Il nostro successo, la nostra felicità e il nostro benessere non sono mai completamente opera nostra. Le decisioni degli altri possono influenzare profondamente il corso della nostra vita: se ammetterci ad una determinata scuola, offrirci un lavoro o concederci un mutuo. Un processo decisionale arbitrario, incoerente o errato solleva quindi serie preoccupazioni perché rischia di limitare la nostra capacità di raggiungere gli obiettivi che ci siamo prefissati e di accedere alle opportunità per le quali siamo qualificati.

Quindi, come possiamo garantire che queste decisioni siano prese nel modo giusto e per le giuste ragioni? Sebbene ci sia molto da apprezzare in regole fisse, applicate in modo coerente, le buone decisioni tengono conto delle prove disponibili. Ci aspettiamo che le decisioni in materia di ammissione, impiego e prestito si basino su fattori rilevanti per l'esito degli interessi.

L'identificazione dei dettagli rilevanti per una decisione potrebbe avvenire in modo informale e senza pensarci troppo: i datori di lavoro potrebbero osservare che le persone che studiano matematica sembrano ottenere risultati particolarmente buoni nel settore finanziario. Ma potrebbero verificare queste osservazioni confrontandole con prove storiche esaminando il grado in cui la specializzazione è correlata al successo sul lavoro. Questo è il lavoro tradizionale della statistica e promette di fornire una base più affidabile per il processo decisionale quantificando il peso da assegnare a determinati dettagli nelle nostre determinazioni.

Un corpo di ricerca ha confrontato l'accuratezza dei modelli statistici con i giudizi degli esseri umani, anche di esperti con anni di esperienza. In molti confronti diretti su compiti fissi, le decisioni basate sui dati sono più accurate di quelle basate sull'intuizione o sulla competenza. Ad esempio, in uno studio del 2002, la sottoscrizione automatizzata di prestiti era più accurata e meno disparata dal punto di vista razziale.¹ Questi risultati sono stati accolti come un modo per garantire che le decisioni ad alto rischio che modellano le nostre possibilità di vita siano accurate ed eque.

L'apprendimento automatico promette di portare una maggiore disciplina nel processo decisionale perché offre la possibilità di scoprire fattori rilevanti per il processo decisionale che gli esseri umani potrebbero trascurare, data la complessità o la sottigliezza delle relazioni nelle prove storiche. Invece di partire da qualche intuizione sulla relazione tra determinati fattori e un risultato di interesse, l'apprendimento automatico ci consente di rinviare la questione della rilevanza ai dati stessi: quali fattori, tra tutti quelli che abbiamo osservato, hanno una relazione statistica con il risultato.

Scoprire modelli nelle prove storiche può essere ancora più potente di quanto ciò sembri suggerire. Scoperte nella visione artificiale, in particolare negli oggetti

riconoscimento: rivela quanto può raggiungere la scoperta di modelli. In questo ambito, l'apprendimento automatico ha contribuito a superare uno strano fatto della cognizione umana: anche se possiamo essere in grado di identificare senza sforzo gli oggetti in una scena, non siamo in grado di specificare l'intero insieme di regole su cui facciamo affidamento per prendere queste decisioni. Non possiamo codificare manualmente un programma che enumera in modo esaustivo tutti i fattori rilevanti che ci permettono di riconoscere gli oggetti da ogni prospettiva possibile o in tutte le loro potenziali configurazioni visive. L'apprendimento automatico mira a risolvere questo problema abbandonando il tentativo di insegnare a un computer attraverso istruzioni esplicite a favore di un processo di apprendimento tramite esempi. Esponendo il computer a molti esempi di immagini contenenti oggetti pre-identificati, speriamo che il computer impari i modelli che distinguono in modo affidabile i diversi oggetti l'uno dall'altro e dagli ambienti in cui appaiono.

Questo può sembrare un risultato notevole, non solo perché ora i computer possono eseguire compiti complessi, ma anche perché le regole per decidere cosa appare in un'immagine sembrano emergere dai dati stessi.

Ma ci sono seri rischi nell'imparare dagli esempi. L'apprendimento non è un processo che consiste semplicemente nel memorizzare degli esempi nella memoria. Si tratta invece di generalizzare a partire dagli esempi: focalizzarsi su quei dettagli che sono caratteristici (ad esempio) dei gatti in generale, non solo sui gatti specifici che appaiono negli esempi. Questo è il processo di induzione: trarre regole generali da esempi specifici – regole che effettivamente tengono conto dei casi passati, ma si applicano anche a casi futuri, non ancora visti. La speranza è che riusciremo a capire in che modo i casi futuri potrebbero essere simili a quelli passati, anche se non sono esattamente gli stessi.

Ciò significa che per generalizzare in modo affidabile da esempi storici a casi futuri è necessario fornire al computer buoni esempi: un numero sufficientemente ampio di esempi per scoprire modelli sottili; una serie di esempi sufficientemente diversificata per mostrare i molti diversi tipi di aspetto che gli oggetti potrebbero assumere; una serie di esempi sufficientemente ben annotati per fornire all'apprendimento automatico una verità fondamentale affidabile; e così via. Pertanto, il processo decisionale basato sull'evidenza è affidabile quanto lo sono le prove su cui si basa, e gli esempi di alta qualità sono di fondamentale importanza per l'apprendimento automatico. Il fatto che l'apprendimento automatico sia "basato sull'evidenza" non garantisce in alcun modo che porterà a decisioni accurate, affidabili o giuste.

Ciò è particolarmente vero quando si utilizza l'apprendimento automatico per modellare il comportamento e le caratteristiche umane. I nostri esempi storici dei risultati rilevanti rifletteranno quasi sempre i pregiudizi storici contro determinati gruppi sociali, gli stereotipi culturali prevalenti e le disuguaglianze demografiche esistenti. E trovare modelli in questi dati significherà spesso replicare queste stesse dinamiche.

Qualcos'altro va perso nel passaggio a un processo decisionale automatizzato e predittivo. I decisorи umani raramente cercano di massimizzare l'accuracy predittiva a tutti i costi; spesso, potrebbero considerare fattori come se gli attributi utilizzati per la previsione siano moralmente rilevanti. Ad esempio, sebbene gli imputati più giovani abbiano statisticamente maggiori probabilità di recidiva, i giudici sono restii a tenerne conto nel decidere la durata della pena, considerando gli imputati più giovani come meno moralmente colpevoli.

Questo è uno dei motivi per essere cauti nei confronti dei confronti che sembrano mostrare la superiorità

del processo decisionale statistico.² È inoltre improbabile che gli esseri umani prendano decisioni ovviamente assurde, ma ciò potrebbe accadere con un processo decisionale automatizzato, forse a causa di dati errati. Queste e molte altre differenze tra il processo decisionale umano e quello automatizzato sono le ragioni per cui i sistemi decisionali che si basano sull'apprendimento automatico potrebbero essere ingiusti.

Scriviamo questo libro mentre l'apprendimento automatico inizia a svolgere un ruolo particolarmente importante nei processi decisionali. Nel sistema di giustizia penale, come accennato in precedenza, agli imputati vengono assegnati punteggi statistici destinati a prevedere il rischio di commettere crimini futuri, e questi punteggi informano le decisioni su cauzione, condanna e libertà condizionale. Nella sfera commerciale, le aziende utilizzano l'apprendimento automatico per analizzare e filtrare i curriculum dei candidati al lavoro. E i metodi statistici sono ovviamente il pane quotidiano della concessione di prestiti, crediti e assicurazioni.

Iniziamo ora a esaminare i rischi in queste e molte altre applicazioni dell'apprendimento automatico e forniamo un'analisi critica di una serie emergente di soluzioni proposte. Vedremo come anche applicazioni ben intenzionate dell'apprendimento automatico potrebbero dare origine a risultati discutibili.

Disparità demografiche

Amazon utilizza un sistema basato sui dati per determinare i quartieri in cui offrire la consegna gratuita in giornata. Un'indagine del 2016 ha rilevato forti disparità nella composizione demografica di questi quartieri: in molte città degli Stati Uniti, i residenti bianchi avevano più del doppio delle probabilità rispetto ai residenti neri di vivere in uno dei quartieri idonei.³ Ora, non conosciamo i

dettagli di come funziona il sistema di Amazon, e in particolare non sappiamo in che misura utilizzi il machine learning. Lo stesso vale per molti altri sistemi riportati dalla stampa. Tuttavia, li utilizzeremo come esempi motivanti in cui un sistema di apprendimento automatico per l'attività in questione mostrerebbe plausibilmente lo stesso comportamento.

Nel capitolo 3 vedremo come rendere matematicamente precisa la nostra intuizione sulle disparità demografiche, e vedremo che esistono molti modi possibili per misurare queste disuguaglianze. La pervasività di tali disparità nelle applicazioni di machine learning è una delle principali preoccupazioni di questo libro.

Quando osserviamo le disparità, ciò non implica che il progettista del sistema intendesse che tali disuguaglianze si verificassero. Guardando oltre le intenzioni, è importante capire quando le disparità osservate possono essere considerate discriminazioni. A loro volta, due domande chiave da porsi sono se le disparità siano giustificate e se siano dannose. Queste domande raramente hanno risposte semplici, ma l'ampia letteratura sulla discriminazione in filosofia e sociologia può aiutarci a ragionare su di esse.

Per capire perché le disparità razziali nel sistema di Amazon potrebbero essere dannose, dobbiamo tenere presente la storia del pregiudizio razziale negli Stati Uniti, la sua relazione con la segregazione e le disparità geografiche e la perpetuazione di tali disuguaglianze nel tempo. Amazon ha sostenuto che il suo sistema era giustificato perché

è stato progettato sulla base di considerazioni di efficienza e costi e quella razza non era un fattore esplicito. Tuttavia, ha l'effetto di offrire diverse opportunità ai consumatori a tassi razzialmente diversi. La preoccupazione è che ciò possa contribuire al perpetuarsi di cicli di disuguaglianza di lunga durata. Se, invece, il sistema fosse risultato parziale rispetto ai CAP che terminano con una cifra dispari, non avrebbe scatenato una simile protesta.

Il termine bias è spesso usato per riferirsi a disparità demografiche nei sistemi algoritmici che sono discutibili per ragioni sociali. In questo libro ridurremo al minimo l'uso di questo senso della parola pregiudizio, poiché discipline e comunità diverse interpretano il termine in modo diverso e questo può creare confusione. Esiste un uso più tradizionale del termine bias nelle statistiche e nell'apprendimento automatico. Supponiamo che le stime di Amazon relative alle date/orari di consegna siano costantemente in anticipo di alcune ore. Questo sarebbe un caso di bias statistico. Uno stimatore statistico si dice distorto se il suo valore atteso o medio differisce dal valore reale che si propone di stimare. Il bias statistico è un concetto fondamentale in statistica ed esiste un ricco insieme di tecniche consolidate per analizzarlo ed evitarlo.

Esistono molte altre misure che quantificano le proprietà statistiche desiderabili di un predittore o di uno stimatore, come precisione, richiamo e calibrazione. Anche questi sono ben compresi; nessuno di essi richiede alcuna conoscenza dei gruppi sociali e sono relativamente semplici da misurare. L'attenzione ai criteri demografici nelle statistiche e nell'apprendimento automatico è una direzione relativamente nuova. Ciò riflette un cambiamento nel modo in cui concettualizziamo i sistemi di apprendimento automatico e le responsabilità di coloro che li costruiscono. Il nostro obiettivo è riflettere fedelmente i dati? Oppure abbiamo l'obbligo di mettere in discussione i dati e di progettare i nostri sistemi in modo che si conformino a qualche nozione di comportamento equo, indipendentemente dal fatto che ciò sia supportato o meno dai dati attualmente a nostra disposizione? Queste prospettive sono spesso in conflitto e la differenza tra loro diventerà più chiara quando approfondiremo le fasi dell'apprendimento automatico .

Il ciclo di apprendimento automatico

Studiamo la pipeline del machine learning e comprendiamo come le disparità demografiche si propagano attraverso di essa. Questo approccio ci consente di intravedere la scatola nera del machine learning e ci preparerà per le analisi più dettagliate nei capitoli successivi. Studiare le fasi del machine learning è fondamentale se si vuole intervenire per ridurre al minimo le disparità.

La figura seguente mostra le fasi di un tipico sistema che produce output utilizzando l'apprendimento automatico. Come ogni diagramma di questo tipo, è una semplificazione, ma è utile per i nostri scopi.

La prima fase è la misurazione, ovvero il processo mediante il quale lo stato del mondo viene ridotto a un insieme di righe, colonne e valori in un set di dati. È un processo disordinato, perché il mondo reale è disordinato. Il termine misurazione è fuorviante, poiché evoca l'immagine di uno scienziato imparziale che registra ciò che osserva, mentre vedremo che richiede decisioni umane soggettive.

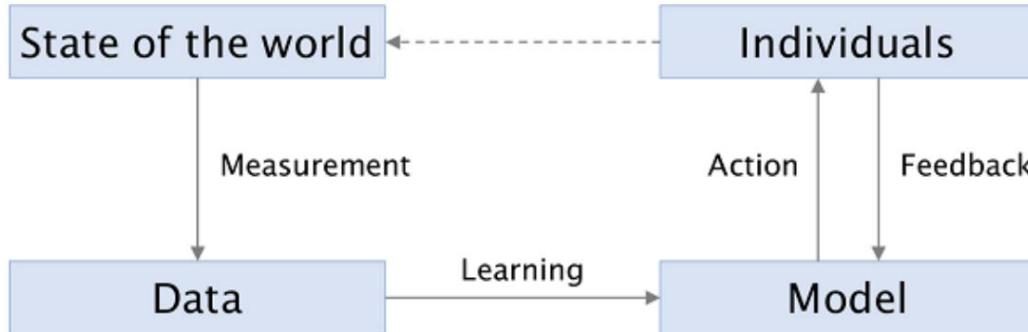


Figura 1.1: Il ciclo di apprendimento automatico

L'"apprendimento" nell'apprendimento automatico si riferisce alla fase successiva, ovvero trasformare i dati in un modello. Un modello riepiloga i modelli nei dati di addestramento; fa generalizzazioni. Un modello può essere addestrato utilizzando l'apprendimento supervisionato tramite un algoritmo come Support Vector Machines o utilizzando l'apprendimento non supervisionato tramite un algoritmo come il clustering k-means. Potrebbe assumere molte forme: un iperpiano o un insieme di regioni nello spazio n-dimensionale, o un insieme di distribuzioni. In genere è rappresentato come un insieme di pesi o parametri.

La fase successiva è l'azione che intraprendiamo in base alle previsioni del modello, che sono applicazioni del modello a input nuovi e invisibili. A proposito, "previsione" è un altro termine fuorviante, anche se a volte implica il tentativo di predire il futuro ("questo paziente è ad alto rischio di cancro?"), a volte no ("questo account di social media è un bot?") ?".

La previsione può assumere la forma di classificazione (determinare se un messaggio di posta elettronica è spam), regressione (assegnando punteggi di rischio agli imputati) o recupero di informazioni (trovando documenti che corrispondono meglio a una query di ricerca). Le azioni in queste tre applicazioni potrebbero essere: depositare l'e-mail nella casella di posta o nella cartella spam dell'utente, decidere se fissare una cauzione per il rilascio cautelare dell'imputato e visualizzare all'utente i risultati della ricerca recuperati. Possono differire notevolmente nel loro significato per l'individuo, ma hanno in comune il fatto che le risposte collettive degli individui a queste decisioni alterano lo stato del mondo, cioè i modelli sottostanti che il sistema mira a modellare.

Alcuni sistemi di machine learning registrano il feedback degli utenti (come gli utenti reagiscono alle azioni) e li utilizzano per perfezionare il modello. Ad esempio, i motori di ricerca tengono traccia di ciò su cui gli utenti fanno clic come segnale隐式 di pertinenza o qualità. Il feedback può verificarsi anche involontariamente o addirittura in modo contraddittorio; questi sono più problematici, come esploreremo più avanti in questo capitolo.

Lo stato della società

In questo libro ci occupiamo delle applicazioni dell'apprendimento automatico che coinvolgono dati sulle persone. In queste applicazioni, i dati di formazione disponibili codificheranno probabilmente le disparità demografiche esistenti nella nostra società. Ad esempio, il

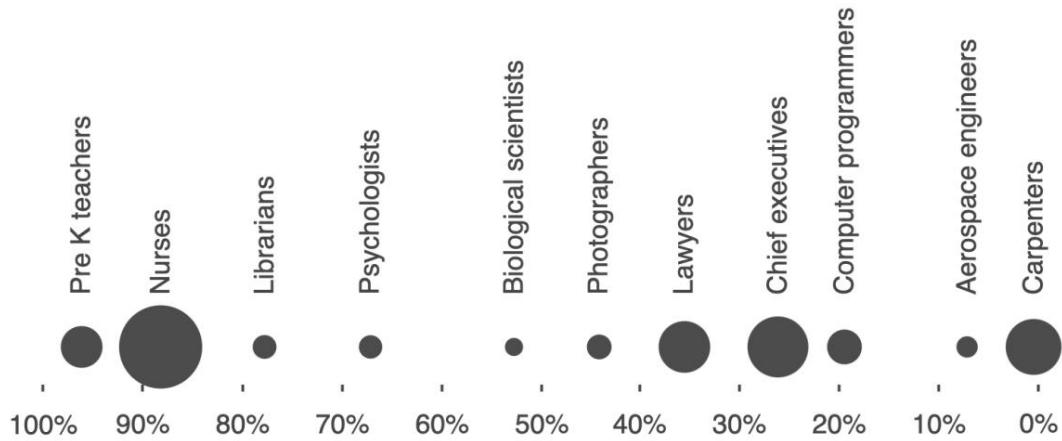


Figura 1.2: Un campione di occupazioni negli Stati Uniti in ordine decrescente in base alla percentuale di donne. L'area della bolla rappresenta il numero di lavoratori.

la figura mostra la ripartizione per genere di un campione di occupazioni negli Stati Uniti, sulla base dei dati diffusi dal Bureau of Labor Statistics per l'anno 2017.

Non sorprende che molte occupazioni presentino forti squilibri di genere. Se stiamo costruendo un sistema di apprendimento automatico che seleziona i candidati per un lavoro, dovremmo essere profondamente consapevoli che questa è la linea di base da cui stiamo iniziando. Ciò non significa necessariamente che i risultati del nostro sistema saranno imprecisi o discriminatori, ma in questo capitolo vedremo come ciò complica le cose.

Perché esistono queste disparità? Ci sono molti fattori che potenzialmente contribuiscono, tra cui una storia di discriminazione esplicita, atteggiamenti impliciti e stereotipi riguardo al genere e differenze nella distribuzione di alcune caratteristiche per genere. Vedremo che anche in assenza di discriminazione esplicita, gli stereotipi possono autoavverarsi e persistere a lungo nella società. Man mano che integriamo l'apprendimento automatico nel processo decisionale, dovremmo fare attenzione a garantire che il machine learning non diventi parte di questo ciclo di feedback.

E le applicazioni che non riguardano le persone? Prendiamo in considerazione "Street Bump", un progetto della città di Boston per raccogliere dati sulle buche. L'app per smartphone rileva automaticamente le buche utilizzando i dati dei sensori dello smartphone e invia i dati alla città. L'infrastruttura sembra un'applicazione comodamente noiosa del processo decisionale basato sui dati, molto lontana dai dilemmi etici di cui abbiamo discusso. Eppure! Kate Crawford sottolinea che i dati riflettono i modelli di possesso di smartphone, che sono più elevati nelle zone più ricche della città rispetto alle aree a basso reddito e alle aree con una grande popolazione anziana.⁴ La lezione qui è che è raro che le applicazioni di machine learning non riguardare le persone. Nel caso di Street Bump, i dati vengono raccolti dalle persone e quindi riflettono le disparità demografiche; inoltre, il motivo per cui siamo interessati a migliorare le infrastrutture in primo luogo è il loro effetto sulla vita delle persone.

Per dimostrare che la maggior parte delle applicazioni di machine learning coinvolgono le persone, abbiamo analizzato Kaggle, una nota piattaforma per le competizioni di data science.

Ci siamo concentrati sui primi 30 concorsi ordinati per importo del premio. In 14 di questi concorsi, abbiamo osservato che il compito è prendere decisioni sugli individui.

Nella maggior parte di questi casi esistono stereotipi o disparità sociali che possono essere perpetuati dall'applicazione dell'apprendimento automatico. Ad esempio, l'attività Automated Essay Scoring5 cerca algoritmi che tentano di far corrispondere i punteggi dei valutatori umani dei saggi degli studenti. Le scelte linguistiche degli studenti sono indicatori dell'appartenenza a un gruppo sociale , ed è noto che i valutatori umani a volte hanno pregiudizi basati su tali fattori.^{6, 7} Pertanto, poiché i valutatori umani devono fornire le etichette originali, i sistemi di valutazione automatizzati rischiano di consacrare qualsiasi modello discriminatorio che sia catturato nei dati di allenamento.

In altri 5 dei 30 concorsi, il compito non richiedeva di prendere decisioni sulle persone, ma le decisioni prese utilizzando il modello avrebbero comunque avuto un impatto diretto sulle persone. Ad esempio, un concorso sponsorizzato dalla società immobiliare Zillow chiede di migliorare l'algoritmo "Zestimate" della società per prevedere i prezzi di vendita delle case. Qualsiasi sistema che preveda il prezzo di vendita futuro di una casa (e pubblicizzi queste previsioni) creerà probabilmente un ciclo di feedback che si autoavvera in cui le case che si prevede avranno prezzi di vendita più bassi scoraggiano i futuri acquirenti, sopprimendo la domanda e abbassando il prezzo di vendita finale.

In 9 dei 30 concorsi, non abbiamo riscontrato un impatto evidente e diretto sulle persone, come nel caso di un concorso sulla previsione della salute degli oceani (ovviamente, anche tali concorsi hanno un impatto indiretto sulle persone, a causa delle azioni che potremmo intraprendere sulla base delle conoscenze acquisite). In due casi, non avevamo informazioni sufficienti per prendere una decisione.

Per riassumere, la società umana è piena di disparità demografiche e i dati sulla formazione probabilmente le rifletteranno. Passeremo ora al processo mediante il quale vengono costruiti i dati di addestramento e vedremo che le cose sono ancora più complicate.

Il problema della misurazione

Il termine misurazione suggerisce un processo semplice, che ricorda una telecamera che registra oggettivamente una scena. In effetti, la misurazione è irta di decisioni soggettive e difficoltà tecniche.

Consideriamo un compito apparentemente semplice: misurare la diversità demografica dei campus universitari. Un articolo del New York Times del 2017 mirava a fare proprio questo ed era intitolato "Anche con l'azione affermativa, i neri e gli ispanici sono più sottorappresentati nelle università più prestigiose rispetto a 35 anni fa".⁸ Gli autori sostengono che il divario tra le matricole nere e ispaniche iscritte e la popolazione nera e ispanica in età universitaria è cresciuta negli ultimi 35 anni. Per supportare la loro affermazione, presentano informazioni demografiche per più di 100 università e college americani dal 1980 al 2015 e mostrano come le percentuali di studenti neri, ispanici, asiatici, bianchi e multirazziali siano cambiate nel corso degli anni. È interessante notare che la categoria multirazziale è stata introdotta solo di recente nel 2008, ma i confronti nell'articolo ignorano l'introduzione di questa nuova categoria.

Quanti studenti che avrebbero potuto selezionare la casella "Bianco" o "Nero" hanno selezionato la casella "Bianco" o "Nero".

invece la scatola "multirazziale"? In che modo ciò potrebbe aver influenzato le percentuali di studenti "bianchi" e "neri" in queste università? Inoltre, la concezione della razza da parte degli individui e della società cambia nel tempo. Una persona con genitori neri e latini sarebbe più propensa a identificarsi come nera nel 2015 rispetto agli anni '80? Il punto è che è impossibile rispondere anche a una domanda apparentemente semplice sulle tendenze della diversità demografica senza fare alcune ipotesi, e illustra le difficoltà di misurazione in un mondo che resiste a cadere ordinatamente in una serie di caselle di controllo. La razza non è una categoria stabile; il modo in cui misuriamo la razza spesso cambia il modo in cui la concepiamo, e il cambiamento delle concezioni di razza può costringerci a modificare ciò che misuriamo.

Per essere chiari, questa situazione è tipica: misurare quasi tutti gli attributi delle persone è altrettanto soggettivo e impegnativo. Anzi, le cose diventano più caotiche quando i ricercatori che si occupano di machine learning devono creare categorie, come spesso accade.

Un'area in cui i professionisti dell'apprendimento automatico devono spesso definire nuove categorie è la definizione della variabile target.⁹ Questo è il risultato che stiamo cercando di prevedere: l'imputato recidiverà se rilasciato su cauzione? Il candidato sarà un buon impiegato se assunto? E così via.

I bias nella definizione della variabile target sono particolarmente critici, perché sicuramente distorcono le previsioni relative al costrutto reale che intendevamo prevedere, come nel caso in cui usiamo gli arresti come misura del crimine, o le vendite come misura del crimine, prestazione lavorativa o GPA come misura del successo accademico. Questo non è necessariamente così con altri attributi. Ma la variabile target è probabilmente la più difficile dal punto di vista della misurazione, perché spesso è un costrutto inventato per gli scopi del problema in questione piuttosto che uno ampiamente compreso e misurato. Ad esempio, l'"affidabilità creditizia" è un costrutto creato nel contesto del problema di come estendere con successo il credito ai consumatori;⁹ non è una proprietà intrinseca che le persone possiedono o non possiedono.

Se la nostra variabile target è l'idea di "buon dipendente", potremmo utilizzare i punteggi di revisione delle prestazioni per quantificarla. Ciò significa che i nostri dati ereditano eventuali bias presenti nelle valutazioni dei manager sui loro report. Un altro esempio: l'uso della visione artificiale per classificare automaticamente l'attrattiva fisica delle persone.^{10, 11} I dati di addestramento consistono nella valutazione umana dell'attrattiva e, non sorprende, tutti questi classificatori hanno mostrato una preferenza per la pelle più chiara.

In alcuni casi potremmo essere in grado di avvicinarci a una definizione più obiettiva di una variabile target, almeno in linea di principio. Ad esempio, nella valutazione del rischio penale, i dati di formazione non riguardano le decisioni dei giudici sulla cauzione, ma piuttosto quelli basati su chi ha effettivamente commesso un reato. Ma c'è almeno un grosso avvertimento: non possiamo davvero misurare chi ha commesso un crimine, quindi usiamo gli arresti come proxy. Ciò significa che i dati sulla formazione contengono distorsioni non dovute ai pregiudizi dei giudici ma a causa di un'attività di polizia discriminatoria. D'altra parte, se la nostra variabile obiettivo fosse se l'imputato si presenta o non si presenta in tribunale per il processo, saremmo in grado di misurarlo direttamente con perfetta precisione. Detto questo, potremmo ancora nutrire preoccupazioni riguardo a un sistema che tratta gli imputati in modo diverso in base alla probabilità prevista di comparizione, dato che alcune ragioni per non comparire sono meno discutibili di altre (cercare di mantenere un lavoro che non consenta il tempo libero rispetto a

cercando di evitare procedimenti

giudiziari).¹² Nelle assunzioni, invece di fare affidamento sulle revisioni delle prestazioni per (ad esempio) un lavoro di vendita, potremmo fare affidamento sul numero di vendite chiuse. Ma si tratta di una misurazione oggettiva o è soggetta ai pregiudizi dei potenziali clienti (che potrebbero rispondere più positivamente a determinati venditori rispetto ad altri) e alle condizioni del posto di lavoro (che potrebbe essere un ambiente ostile per alcuni, ma non per altri)?

In alcune applicazioni, i ricercatori ripropongono uno schema di classificazione esistente per definire la variabile target anziché crearne uno da zero. Ad esempio, un sistema di riconoscimento degli oggetti può essere creato addestrando un classificatore su ImageNet, un database di immagini organizzato in una gerarchia di concetti.¹³ La gerarchia di ImageNet deriva da Wordnet, un database di parole, categorie e le relazioni tra loro.¹⁴ gli autori a loro volta importarono gli elenchi di parole da una serie di fonti più antiche, come i thesauri. Di conseguenza, le categorie di WordNet (e ImageNet) contengono numerose parole e associazioni antiquate, come occupazioni che non esistono più e associazioni di genere

stereotipate.¹⁵ Pensiamo alla tecnologia in rapido cambiamento e alla società lenta nell'adattarsi, ma almeno in questo caso, lo schema di categorizzazione alla base di gran parte della tecnologia di apprendimento automatico di oggi è rimasto congelato nel tempo mentre le norme sociali sono cambiate.

Il nostro esempio preferito di bias di misurazione ha a che fare con le telecamere, a cui abbiamo fatto riferimento all'inizio della sezione come esempio di osservazione e registrazione imparziale. Ma lo sono?

Il mondo visivo ha una larghezza di banda essenzialmente infinita rispetto a ciò che può essere catturato dalle fotocamere, sia a pellicola che digitali, il che significa che la tecnologia fotografica implica una serie di scelte su ciò che è rilevante e ciò che non lo è, e trasformazioni dei dati catturati basati su quelle scelte. Sia le fotocamere a pellicola che quelle digitali sono state storicamente più abili nel fotografare individui dalla pelle più chiara.¹⁶ Uno dei motivi sono le impostazioni predefinite, come il bilanciamento del colore, che sono state ottimizzate per le tonalità della pelle più chiare. Un'altra ragione più profonda è la "gamma dinamica" limitata delle fotocamere, che rende difficile catturare toni più luminosi e più scuri nella stessa immagine. La situazione iniziò a cambiare negli anni '70, in parte a causa delle lamentele delle aziende di mobili e di cioccolato sulla difficoltà di catturare fotograficamente i dettagli rispettivamente dei mobili e del cioccolato!

Un altro impulso è venuto dalla crescente diversità degli argomenti televisivi tempo.

Quando passiamo dalle singole immagini ai set di dati di immagini, introduciamo un altro livello di potenziali pregiudizi. Consideriamo i set di dati di immagini utilizzati per addestrare gli odierni sistemi di visione artificiale per attività come il riconoscimento degli oggetti. Se questi set di dati fossero campioni rappresentativi di un mondo visivo sottostante, potremmo aspettarci che un sistema di visione artificiale addestrato su uno di questi set di dati funzionerebbe bene su un altro set di dati. Ma in realtà, osserviamo un forte calo di accuratezza quando ci alleniamo e testiamo su set di dati diversi.¹⁷ Ciò dimostra che questi set di dati sono distorti l'uno rispetto all'altro in senso statistico, ed è un buon punto di partenza per indagare se questi pregiudizi includano fattori culturali. stereotipi.

Non ci sono solo brutte notizie: l'apprendimento automatico può infatti aiutare a mitigare le misure

pregiudizi mentali. Tornando alla questione della gamma dinamica delle fotocamere, le tecniche computazionali, compreso l'apprendimento automatico, permettono di migliorare la rappresentazione dei toni nelle immagini.^{18, 19, 20} Un altro esempio viene dalla medicina: diagnosi e cure sono talvolta personalizzate in base alla razza. Ma si scopre che la razza viene utilizzata come indicatore grezzo di ascendenza e genetica, e talvolta di fattori ambientali e comportamentali.^{21, 22} Se riusciamo a misurare i fattori che sono rilevanti dal punto di vista medico e incorporarli, al posto della razza, in modelli statistici della risposta alle malattie e ai farmaci, possiamo aumentare l'accuratezza delle diagnosi e dei trattamenti mitigando al contempo le disparità razziali.

Per riassumere, la misurazione implica la definizione delle variabili di interesse, il processo di interazione con il mondo reale e la trasformazione delle osservazioni in numeri, quindi la raccolta effettiva dei dati. Spesso i professionisti del machine learning non pensano a questi passaggi, perché qualcun altro ha già fatto quelle cose. Eppure è fondamentale capire la provenienza dei dati. Anche se qualcun altro ha raccolto i dati, è quasi sempre troppo complicato da gestire per gli algoritmi, da qui il temuto passaggio di "pulizia dei dati". Ma il disordine del mondo reale non è solo un fastidio da affrontare con la pulizia. È una manifestazione di un mondo diversificato in cui le persone non rientrano perfettamente nelle categorie. Essere disattenti a queste sfumature può danneggiare particolarmente le popolazioni emarginate.

Dai dati ai modelli

Abbiamo visto che i dati di addestramento riflettono le disparità, le distorsioni e i pregiudizi del mondo reale e del processo di misurazione. Ciò porta a una domanda ovvia: quando impariamo un modello da tali dati, queste disparità vengono preservate, mitigate o esacerbate?

I modelli predittivi addestrati con metodi di apprendimento supervisionato sono spesso efficaci nella calibrazione: garantiscono che la previsione del modello sussuma tutte le caratteristiche dei dati allo scopo di prevedere il risultato. Ma la calibrazione significa anche che, per impostazione predefinita, dovremmo aspettarci che i nostri modelli riflettano fedelmente le disparità riscontrate nei dati di input.

Ecco un altro modo di pensarci. Alcuni modelli nei dati di addestramento (il fumo è associato al cancro) rappresentano la conoscenza che desideriamo estrarre utilizzando l'apprendimento automatico, mentre altri modelli (le ragazze come il rosa e i ragazzi come il blu) rappresentano gli stereotipi che potremmo voler evitare di apprendere. Ma gli algoritmi di apprendimento non hanno un modo generale per distinguere tra questi due tipi di modelli, perché sono il risultato di norme sociali e giudizi morali. In assenza di un intervento specifico, l'apprendimento automatico estrarrà gli stereotipi, compresi quelli errati e dannosi, nello stesso modo in cui estrae la conoscenza.

Un esempio significativo di ciò viene dalla traduzione automatica. Lo screenshot a destra mostra il risultato della traduzione di frasi dall'inglese al turco e viceversa.²³ Le stesse traduzioni stereotipate risultano per molte coppie di lingue e altre parole professionali in tutti i motori di traduzione che abbiamo testato. È facile capire perché. Il turco ha pronomi neutri rispetto al genere e quando si traduce un pronome del genere

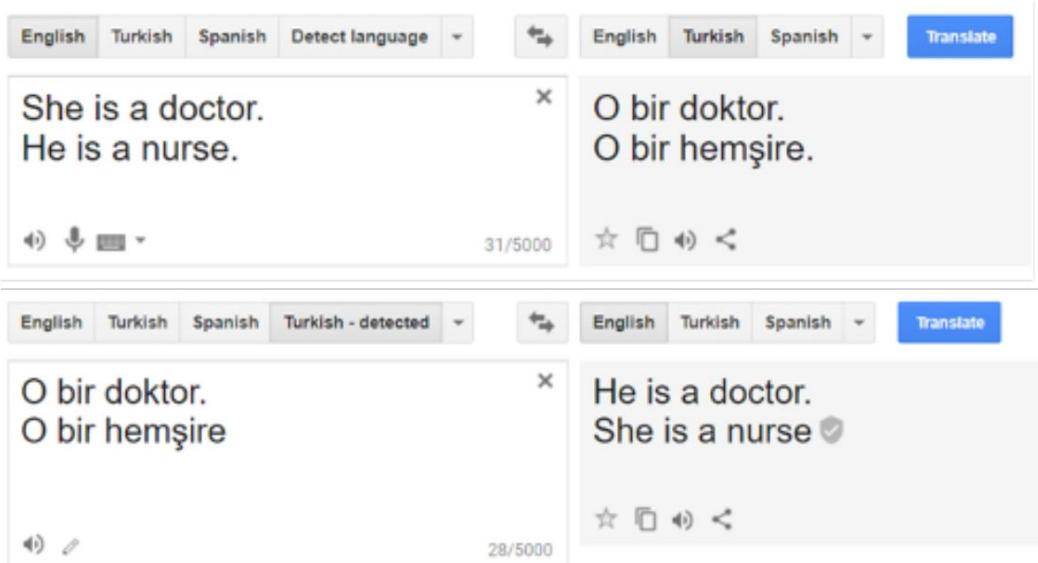


Figura 1.3: Tradurre dall'inglese al turco e poi di nuovo all'inglese introduce stereotipi di genere.

all'inglese, il sistema sceglie la frase che meglio corrisponde alle statistiche del training set (che in genere è un ampio corpus minimamente curato di testo storico e testo trovato sul web).

Quando costruiamo un modello statistico del linguaggio a partire da tale testo, dovremmo aspettarci che le associazioni di genere delle parole relative all'occupazione rispecchino approssimativamente le statistiche sul lavoro del mondo reale. Inoltre, a causa della distorsione del maschio come norma²⁴ (l'uso di pronomi maschili quando il genere è sconosciuto) dovremmo aspettarci che le traduzioni favoriscano i pronomi maschili. Si scopre che quando ripetiamo l'esperimento con dozzine di parole relative all'occupazione, questi due fattori – le statistiche sul lavoro e la propensione al maschio come norma – insieme prevedono quasi perfettamente quale pronome verrà restituito.²⁵

Ecco una risposta allettante all'osservazione che i modelli riflettono errori nei dati. Supponiamo di costruire un modello per il punteggio del curriculum per un lavoro di programmazione. E se semplicemente escludessimo il genere dai dati? È una risposta sufficiente alle preoccupazioni sulla discriminazione di genere? Sfortunatamente, non è così semplice, a causa del problema dei proxy⁹ o delle codifiche ridondanti,²⁵ come discuteremo nel capitolo 3. Nei dati sono presenti numerosi altri attributi che potrebbero essere correlati al genere. Ad esempio, nella nostra società, l'età in cui qualcuno inizia a programmare è correlata al genere. Ciò illustra il motivo per cui non possiamo semplicemente sbarazzarci dei proxy: potrebbero essere veramente rilevanti per la decisione in questione. Da quanto tempo qualcuno programma è un fattore che ci fornisce informazioni preziose sulla sua idoneità per un lavoro di programmazione, ma riflette anche la realtà degli stereotipi di genere.

Un altro motivo comune per cui il machine learning potrebbe funzionare peggio per alcuni gruppi rispetto ad altri è la disparità nella dimensione del campione. Se costruiamo il nostro set di addestramento campionando in modo uniforme i dati di addestramento, per definizione avremo meno dati

punti sulle minoranze. Naturalmente, l'apprendimento automatico funziona meglio quando ci sono più dati, quindi funzionerà meno bene per i membri dei gruppi minoritari, presupponendo che i membri dei gruppi maggioritari e minoritari siano sistematicamente diversi in termini di compito di previsione.²⁵ Peggio

ancora, in molti contesti le minoranze i gruppi sono sottorappresentati rispetto alle statistiche sulla popolazione. Ad esempio, i gruppi minoritari sono sottorappresentati nel settore tecnologico. Gruppi diversi potrebbero anche adottare la tecnologia a ritmi diversi, il che potrebbe distorcere i set di dati assemblati dai social media. Se i set di formazione venissero estratti da questi contesti non rappresentativi, ci sarebbero ancora meno punti di formazione da parte di individui appartenenti a minoranze.

Quando sviluppiamo modelli di machine learning, in genere ne testiamo solo l' accuratezza complessiva; quindi una statistica del “5% di errore” potrebbe nascondere il fatto che un modello ha prestazioni pessime per un gruppo minoritario. Segnalare i tassi di accuratezza per gruppo ci aiuterà ad avvisarci di problemi come nell'esempio precedente. Nel Capitolo 3 esamineremo i parametri che quantificano la disparità del tasso di errore tra i gruppi.

Esiste un'applicazione dell'apprendimento automatico in cui troviamo tassi di errore particolarmente elevati per i gruppi minoritari: il rilevamento delle anomalie. Questa è l'idea di individuare comportamenti che si discostano dalla norma come prova di abuso contro un sistema. Un buon esempio è la controversia Nymwars, in cui Google, Facebook e altre società tecnologiche miravano a bloccare gli utenti che utilizzavano contenuti non comuni (quindi presumibilmente falsi) nomi.

Supponiamo inoltre che in alcune culture la maggior parte delle persone riceva nomi da un piccolo insieme di nomi, mentre in altre culture i nomi potrebbero essere più diversi e potrebbe essere comune che i nomi siano univoci. Per gli utenti di quest'ultima cultura, è più probabile che un nome popolare sia falso. In altre parole, la stessa caratteristica che costituisce una prova a favore di una previsione in un gruppo potrebbe costituire una prova contro la previsione per un altro gruppo.²⁵ Se non stiamo attenti, gli algoritmi di apprendimento generalizzeranno in base alla cultura maggioritaria, portando a un alto tasso di errore per i gruppi

minoritari. Il tentativo di evitare ciò rendendo il modello più complesso si scontra con un problema diverso: l'adattamento eccessivo ai dati di addestramento, ovvero la raccolta di modelli che emergono a causa del rumore casuale piuttosto che delle differenze reali. Un modo per evitare ciò è modellare esplicitamente le differenze tra i gruppi, sebbene vi siano sfide sia tecniche che etiche ad esso associate.

Le insidie dell'azione

Qualsiasi vero sistema di apprendimento automatico cerca di apportare qualche cambiamento nel mondo. Per comprenderne gli effetti, quindi, dobbiamo considerarlo nel contesto del più ampio sistema socio-tecnico in cui è incorporato.

Nel Capitolo 3 vedremo che se un modello è calibrato (cattura fedelmente i modelli nei dati sottostanti) le previsioni fatte utilizzando quel modello avranno inevitabilmente tassi di errore disparati per gruppi diversi, se tali gruppi hanno tassi di base diversi, cioè , tassi di risultati positivi o negativi. In altre parole, comprensione

le proprietà di una previsione richiedono la comprensione non solo del modello, ma anche delle differenze di popolazione tra i gruppi a cui vengono applicate le previsioni.

Inoltre, le caratteristiche della popolazione possono cambiare nel tempo; questo è un noto fenomeno di apprendimento automatico noto come deriva. Se le sottopopolazioni cambiano in modo diverso nel tempo, ma il modello non viene riqualificato, ciò può introdurre disparità. Un ulteriore problema: il fatto che le disparità siano discutibili o meno può variare da una cultura all'altra e può cambiare nel tempo con l'evoluzione delle norme sociali.

Quando le persone sono soggette a decisioni automatizzate, la loro percezione di tali decisioni dipende non solo dai risultati ma anche dal processo decisionale. Un processo decisionale etico potrebbe richiedere, tra le altre cose, la capacità di spiegare una previsione o una decisione, cosa che potrebbe non essere fattibile con i modelli a scatola nera.

Una delle principali limitazioni dell'apprendimento automatico è che rivela solo correlazioni, ma spesso utilizziamo le sue previsioni come se rivelassero una causalità. Questa è una fonte persistente di problemi. Ad esempio, uno dei primi sistemi di apprendimento automatico nel settore sanitario ha imparato la regola apparentemente priva di senso secondo cui i pazienti con asma avevano un rischio inferiore di sviluppare la polmonite. Questo era un modello vero nei dati, ma la ragione probabile era che i pazienti asmatici avevano maggiori probabilità di ricevere cure ospedaliere.²⁶ Quindi non è valido utilizzare la previsione per decidere se ricoverare o meno un paziente. Discuteremo la causalità nel capitolo 5.

Un altro modo di vedere questo esempio è che la previsione influenza il risultato (a causa delle azioni intraprese sulla base della previsione) e quindi si invalida. Lo stesso principio si riscontra anche nell'uso dell'apprendimento automatico per prevedere la congestione del traffico: se un numero sufficiente di persone sceglie il proprio percorso in base alla previsione, allora il percorso che si prevede sarà libero sarà infatti congestionato. L'effetto può funzionare anche nella direzione opposta: la previsione potrebbe rafforzare il risultato, dando luogo a cicli di feedback. Per capire meglio come, parliamo della fase finale del nostro ciclo: il feedback.

Feedback e cicli di feedback

Molti sistemi ricevono feedback quando fanno previsioni. Quando un motore di ricerca fornisce risultati, in genere registra i collegamenti su cui l'utente fa clic e il tempo che l'utente trascorre su quelle pagine e li tratta come segnali impliciti su quali risultati sono risultati più pertinenti. Quando un sito Web di condivisione video consiglia un video, utilizza il feedback police su/giù come segnale esplicito. Tale feedback viene utilizzato per perfezionare il modello.

Ma il feedback è difficile da interpretare correttamente. Se un utente ha cliccato sul primo collegamento in una pagina dei risultati di ricerca, è semplicemente perché era il primo o perché era effettivamente il più pertinente? Anche questo è il caso dell'azione (l'ordinamento dei risultati della ricerca) che influenza il risultato (il/i collegamento/i su cui l'utente fa clic). Questa è un'area attiva di ricerca; esistono tecniche che mirano a imparare accuratamente da questo tipo di feedback distorto.²⁷

Le distorsioni nel feedback potrebbero anche riflettere pregiudizi culturali, che ovviamente sono molto più difficili da caratterizzare rispetto agli effetti dell'ordinamento dei risultati di ricerca. Ad esempio, i clic sugli annunci mirati visualizzati accanto ai risultati di ricerca potrebbero riflettere stereotipi di genere e razziali. C'è un noto studio di Latanya Sweeney che suggerisce questo: le ricerche su Google per nomi dal suono nero come "Latanya Farrell" avevano molte più probabilità di risultati in annunci di documenti di arresto ("Latanya Farrell, Arrested?") rispetto alle ricerche di Nomi dal suono bianco ("Kristen Haring").²⁸ Una potenziale spiegazione è che è più probabile che gli utenti facciano clic su annunci conformi agli stereotipi e che il sistema pubblicitario sia ottimizzato per massimizzare i clic.

In altre parole, anche il feedback integrato nei sistemi può portare a pregiudizi imprevisti o indesiderati. Ma oltre a ciò, ci sono molti modi inattesi in cui potrebbero verificarsi feedback, e questi sono più dannosi e più difficili da controllare.

Diamo un'occhiata a tre.

Previsioni che si autoavverano. Supponiamo che un sistema di polizia predittiva determini che alcune aree di una città siano ad alto rischio di criminalità. In tali aree potrebbero essere impiegati più agenti di polizia. In alternativa, gli agenti in aree che si prevede siano ad alto rischio potrebbero abbassare sottilmente la loro soglia per fermare, perquisire o arrestare le persone, forse anche inconsciamente. In ogni caso, la previsione sembrerà convalidata, anche se fosse stata fatta esclusivamente sulla base di dati distorti.

Ecco un altro esempio di come agire in base a una previsione può cambiare il risultato. Negli Stati Uniti, alcuni imputati criminali vengono rilasciati prima del processo, mentre per altri viene fissata una cauzione come precondizione per il rilascio. Molti imputati non sono in grado di pagare la cauzione. Il rilascio o la detenzione influiscono sull'esito del caso? Forse gli imputati detenuti subiscono maggiori pressioni affinché si dichiarino colpevoli. In ogni caso, come si potrebbe verificare l'impatto causale della detenzione senza fare un esperimento? Curiosamente, possiamo trarre vantaggio da uno pseudo-esperimento, vale a dire che agli imputati viene assegnata la cauzione ai giudici in modo quasi casuale, e alcuni giudici sono più severi di altri. Pertanto, la custodia cautelare è parzialmente casuale, in modo quantificabile. Gli studi che utilizzano questa tecnica hanno confermato che la detenzione provoca effettivamente un aumento della probabilità di una condanna.²⁹ Se la cauzione fosse fissata sulla base di previsioni di rischio, sia umane che algoritmiche, e ne valutassimo l'efficacia esaminando gli esiti dei casi, vedremmo un'autodeterminazione. effetto appagante.

Previsioni che influenzano il training set. Continuando questo esempio, l'attività di polizia predittiva porterà ad arresti, le cui registrazioni potrebbero essere aggiunte al set di addestramento dell'algoritmo. Queste aree potrebbero quindi continuare a sembrare ad alto rischio di criminalità, e forse anche altre aree con una composizione demografica simile, a seconda dell'insieme di caratteristiche utilizzate per le previsioni. Le disparità potrebbero addirittura aggravarsi nel tempo.

Un articolo del 2016 di Lum e Isaac ha analizzato un algoritmo di polizia predittiva di PredPol. Questo è uno dei pochi algoritmi di polizia predittiva ad essere pubblicato in una rivista peer-reviewed, per il quale l'azienda merita un elogio. Applicando l'algoritmo ai dati derivati dai registri della polizia di Oakland, gli autori hanno scoperto che i neri sarebbero presi di mira per la polizia predittiva dei crimini legati alla droga con un tasso circa doppio rispetto ai bianchi, anche se i due gruppi hanno tassi di consumo di droga più o meno uguali.³⁰ La loro simulazione ha mostrato che questo pregiudizio iniziale sarebbe stato

amplificato da un ciclo di feedback, con la polizia sempre più concentrata su aree mirate. Questo nonostante il fatto che l'algoritmo PredPol non tenga esplicitamente conto dei dati demografici.

Un documento successivo si è basato su questa idea e ha mostrato matematicamente come si verificano i cicli di feedback quando i dati scoperti sulla base delle previsioni vengono utilizzati per aggiornare il modello.³¹ Il documento mostra anche come modificare il modello per evitare cicli di feedback in un ambiente simulato: quantificando quanto sorprendente sia l'osservazione di un crimine data dalle previsioni e aggiornando il modello solo in risposta alla sorpresa eventi.

Previsioni che riguardano il fenomeno e la società in generale. Le attività di polizia pregiudizievoli su larga scala, algoritmiche o meno, influenzano la società nel tempo, contribuendo al ciclo di povertà e criminalità. Si tratta di una tesi ben battuta, e nel capitolo 8 esamineremo brevemente la letteratura sociologica sulla diseguaglianza durevole e sulla persistenza degli stereotipi .

Ricordiamoci che utilizziamo l'apprendimento automatico in modo da poter agire in base alle sue previsioni. È difficile, anche concettualmente, eliminare gli effetti delle previsioni sui risultati, sui futuri set di formazione, sui fenomeni stessi o sulla società in generale.

Quanto più il machine learning diventa centrale nelle nostre vite, tanto più forte sarà questo effetto.

Tornando all'esempio di un motore di ricerca, nel breve termine potrebbe essere possibile estrarre un segnale imparziale dai clic degli utenti, ma nel lungo periodo, i risultati restituiti più spesso verranno collegati e quindi si classificheranno più in alto. Come effetto collaterale del raggiungimento del suo scopo di recuperare informazioni rilevanti, un motore di ricerca cambierà necessariamente proprio ciò che mira a misurare, ordinare e classificare. Allo stesso modo, la maggior parte dei sistemi di apprendimento automatico influenzano i fenomeni che prevedono. Questo è il motivo per cui abbiamo descritto il processo di machine learning come un ciclo.

Nel corso di questo libro impareremo metodi per mitigare i pregiudizi sociali nell'apprendimento automatico, ma dovremmo tenere presente che ci sono limiti fondamentali a ciò che possiamo ottenere, soprattutto se consideriamo l'apprendimento automatico come un sistema socio-tecnico invece che un sistema di apprendimento automatico. astrazione matematica. Il modello da manuale secondo cui i dati di addestramento e test sono indipendenti e distribuiti in modo identico è una semplificazione e potrebbe essere irrealizzabile nella pratica.

Diventare concreto con un esempio di giocattolo

Consideriamo ora un contesto concreto, anche se si tratta di un problema giocattolo, per illustrare molte delle idee discusse finora, e alcune nuove.

Diciamo che fai parte di un comitato di assunzione e prendi decisioni basate solo su due attributi di ciascun candidato: il GPA del college e il punteggio del colloquio (abbiamo detto che è un problema di giocattoli!). Lo formuliamo come un problema di apprendimento automatico: il compito è utilizzare queste due variabili per prevedere una certa misura della "qualità" di un candidato. Ad esempio, potrebbe basarsi sul punteggio medio di revisione delle prestazioni dopo due anni presso l'azienda. Supponiamo di disporre di dati di candidati precedenti che ci consentano di addestrare un modello per prevedere i punteggi delle prestazioni in base al GPA e punteggio dell'intervista.

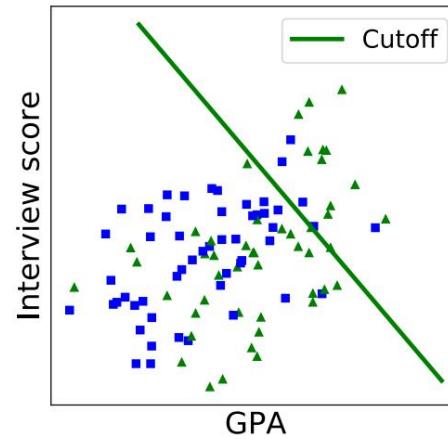


Figura 1.4: Esempio di giocattolo: un classificatore delle assunzioni che prevede la prestazione lavorativa (non mostrata) in base al GPA e al punteggio del colloquio, quindi applica un limite.

Ovviamente, questa è una formulazione riduttiva: presupponiamo che il valore di un candidato possa essere ridotto a un singolo numero e che sappiamo come misurare quel numero. Questa è una critica valida e si applica alla maggior parte delle applicazioni odierne del processo decisionale basato sui dati. Ma ha un grande vantaggio: una volta formulata la decisione come un problema di previsione, i metodi statistici tendono a fare meglio degli esseri umani, anche degli esperti del settore con anni di formazione, nel prendere decisioni basate su predittori rumorosi.

Data questa formulazione, la cosa più semplice che possiamo fare è utilizzare la regressione lineare per prevedere la valutazione media delle prestazioni lavorative dalle due variabili osservate, quindi utilizzare un limite basato sul numero di candidati che vogliamo assumere. La figura sopra mostra come potrebbe apparire. In realtà non è necessario che le variabili in esame soddisfino una relazione lineare, suggerendo quindi l'utilizzo di un modello non lineare, che evitiamo per semplicità.

Come puoi vedere nella figura, i nostri candidati rientrano in due gruppi demografici, rappresentati da triangoli e quadrati. Questa categorizzazione binaria è una semplificazione ai fini del nostro esperimento mentale. Ma quando si costruiscono sistemi reali, imporre rigide categorie di persone può essere eticamente discutibile.

Tieni presente che il classificatore non tiene conto del gruppo a cui apparteneva un candidato. Ciò significa che il classificatore è giusto? Potremmo sperare che, basandosi sull'idea di equità come cecità, sia simboleggiato dall'icona di Lady Justice che indossa una benda. Da questo punto di vista, un modello imparziale, che non utilizzi l'appartenenza al gruppo nella regressione, è giusto; un modello che attribuisce punteggi diversi a membri altrimenti identici di gruppi diversi è discriminatorio.

Rimanderemo una comprensione più approfondita di cosa significhi equità ai capitoli successivi, quindi poniamo una domanda più semplice: i candidati dei due gruppi hanno la stessa probabilità di essere classificati positivamente? La risposta è no: è più probabile che lo siano i triangoli

selezionati rispetto ai quadrati. Questo perché i dati sono uno specchio sociale; le etichette di "verità fondamentale" che prevediamo – valutazioni delle prestazioni lavorative – sono sistematicamente inferiori per i quadrati rispetto ai triangoli.

Ci sono molte possibili ragioni per questa disparità. In primo luogo, i manager che valutano le prestazioni dei dipendenti potrebbero discriminare un gruppo. Oppure il posto di lavoro nel suo complesso potrebbe essere meno accogliente per un gruppo, impedendogli di raggiungere il proprio potenziale e portando a prestazioni inferiori. In alternativa, la disparità potrebbe avere origine prima che i candidati fossero assunti. Ad esempio, potrebbe derivare dalle disparità negli istituti scolastici frequentati dai due gruppi. Oppure potrebbero esserci differenze intrinseche tra loro. Naturalmente, potrebbe essere una combinazione di questi fattori. Non possiamo dire dai nostri dati quanta parte della disparità sia attribuibile a questi diversi fattori. In generale, tale determinazione è metodologicamente difficile e

richiede un ragionamento causale.³² Per ora, supponiamo di avere prove che il livello di disparità demografica prodotto dalla nostra procedura di selezione sia ingiustificato e siamo interessati a intervenire per ridurlo. Come potremmo farlo? Osserviamo che il GPA è correlato all'attributo demografico: è un proxy. Forse potremmo semplicemente omettere quella variabile come predittore? Sfortunatamente, comprometteremmo anche la precisione del nostro modello. Nei set di dati reali, la maggior parte degli attributi tende ad essere proxy delle variabili demografiche e eliminarli potrebbe non essere utile.

Un altro approccio approssimativo consiste nello scegliere diverse soglie di selezione in modo che i candidati di entrambi i gruppi abbiano la stessa probabilità di essere assunti. Oppure potremmo mitigare la disparità demografica invece di eliminarla, diminuendo la differenza tra i limiti.

Dati i dati disponibili, non esiste un modo matematico per sapere quali limiti scegliere. In alcune situazioni esiste una base legale: ad esempio, le linee guida della Commissione statunitense per le pari opportunità di lavoro affermano che se la probabilità di selezione per due gruppi differisce di oltre il 20%, potrebbe costituire un impatto sufficientemente disparato per avviare una causa. Ma un impatto disparato da solo non è illegale; la disparità deve essere ingiustificata o evitabile affinché i tribunali possano accertare la responsabilità. Anche queste linee guida quantitative non forniscono risposte facili o linee chiare.

In ogni caso, l'approccio basato sulla scelta di soglie diverse per mitigare le disparità sembra insoddisfacente, perché è rozzo e utilizza l'attributo di gruppo come unico criterio di ridistribuzione. Non tiene conto delle ragioni di fondo per cui due candidati con gli stessi attributi osservabili (ad eccezione dell'appartenenza al gruppo) potrebbero meritare un trattamento diverso.

Ma ci sono altri possibili interventi e ne discuteremo uno. Per motivarci, facciamo un passo indietro e chiediamoci perché l'azienda vuole ridurre la disparità demografica nelle assunzioni.

Una risposta è radicata nella giustizia verso gli individui e gli specifici gruppi sociali a cui appartengono. Ma una risposta diversa deriva dagli interessi egoistici dell'azienda: team diversi lavorano meglio.^{33, 34} Da questa prospettiva, aumentare la diversità del gruppo assunto andrebbe a beneficio dell'azienda e di tutti i membri del gruppo. Per analogia, scegliere 11 portieri, anche se individualmente eccellenti, equivarrebbe a

povera squadra di calcio.

Come rendiamo operativa la diversità in un compito di selezione? Se avessimo una funzione di distanza tra coppie di candidati, potremmo misurare la distanza media tra i candidati selezionati. Come uomo di paglia, supponiamo di utilizzare la distanza euclidea basata sul GPA e sul punteggio dell'intervista. Se incorporassimo un tale criterio di diversità nella funzione obiettivo, si otterebbe un modello in cui il GPA ha un peso inferiore. Questa tecnica non considera esplicitamente l'appartenenza al gruppo. Piuttosto, come effetto collaterale dell'insistere sulla diversità degli altri attributi osservabili, si migliora anche la diversità demografica. Tuttavia, un'applicazione imprudente di tale intervento può facilmente andare storta: ad esempio, il modello potrebbe dare peso ad attributi che sono completamente irrilevanti per il compito.

Più in generale, ci sono molti possibili interventi algoritmici oltre alla scelta di soglie diverse per gruppi diversi. In particolare, l'idea di una funzione di somiglianza tra coppie di individui è potente e vedremo altri interventi che ne fanno uso. Ma trovare una funzione di somiglianza adeguata nella pratica non è facile: potrebbe non essere chiaro quali attributi siano rilevanti, come ponderarli e come gestire le correlazioni tra gli attributi.

Giustizia oltre il giusto processo decisionale

La preoccupazione principale di questo libro sono le disparità di gruppo nel processo decisionale. Ma gli obblighi etici non si limitano ad affrontare queste disparità. Decisioni equamente prese in circostanze ingiuste possono fare ben poco per migliorare la vita delle persone. In molti casi, non possiamo raggiungere alcuna ragionevole nozione di equità semplicemente modificando il processo decisionale; dobbiamo cambiare le condizioni in cui vengono prese queste decisioni. In altri casi, lo scopo stesso del sistema potrebbe essere oppressivo e dovremmo chiederci se sia opportuno o meno attuarlo.

Inoltre, i sistemi decisionali non sono gli unici luoghi in cui viene utilizzato l'apprendimento automatico che può danneggiare le persone: ad esempio, anche gli algoritmi di ricerca e raccomandazione online destano preoccupazione, anche se non prendono decisioni sulle persone. Discutiamo brevemente queste domande più ampie.

Interventi mirati alle disuguaglianze sottostanti

Torniamo all'esempio di assunzione di cui sopra. Quando utilizziamo l'apprendimento automatico per fare previsioni su come qualcuno potrebbe comportarsi in un posto di lavoro o in un'occupazione specifica, tendiamo a trattare l'ambiente che le persone si troveranno ad affrontare in questi ruoli come una costante e a chiederci come varieranno le prestazioni delle persone in base alle loro caratteristiche osservabili. In altre parole, trattiamo lo stato attuale del mondo come un dato di fatto, lasciando a noi la scelta della persona che si comporterà meglio in queste circostanze. Questo approccio rischia di trascurare i cambiamenti più fondamentali che potremmo apportare al luogo di lavoro (cultura, politiche favorevoli alla famiglia, formazione sul posto di lavoro) che potrebbero renderlo un ambiente più accogliente e produttivo per le persone che non hanno prosperato nelle condizioni precedenti.³⁵

La tendenza nel lavoro sull'equità nell'apprendimento automatico è quella di chiedersi se un datore di lavoro stia utilizzando un processo di selezione equo, anche se potremmo avere l'opportunità di intervenire nelle dinamiche del posto di lavoro che effettivamente tengono conto delle differenze nei risultati previsti in termini di razza, genere , disabilità e altre caratteristiche.³⁶

Possiamo imparare molto dal cosiddetto modello sociale della disabilità, che vede una differenza prevista nella capacità di una persona con disabilità di eccellere sul lavoro come il risultato di una mancanza di sistemazioni adeguate (un ambiente accessibile posto di lavoro, attrezzature necessarie, modalità di lavoro flessibili) piuttosto che qualsiasi capacità intrinseca della persona. Una persona è disabile solo nel senso che non abbiamo costruito ambienti fisici o adottato politiche adeguate per garantirne la pari partecipazione.

Lo stesso potrebbe valere per le persone con altre caratteristiche, e i soli cambiamenti al processo di selezione non ci aiuteranno ad affrontare la fondamentale ingiustizia delle condizioni che impediscono ad alcune persone di contribuire con la stessa efficacia di altre. Esamineremo queste domande nel capitolo 8.

Potrebbe non essere affatto etico implementare un sistema decisionale automatizzato se le condizioni sottostanti sono ingiuste e il sistema automatizzato servirebbe solo a reificarlo. Oppure un sistema può essere mal concepito e il suo scopo può essere ingiusto, anche se funzionasse perfettamente e funzionasse altrettanto bene per tutti.

La questione di quali sistemi automatizzati debbano essere implementati non dovrebbe essere lasciata alla logica (e ai capricci) del mercato. Ad esempio, potremmo voler regolamentare l'accesso della polizia al riconoscimento facciale. I nostri diritti civili – libertà, movimento e associazione – sono minacciati da queste tecnologie sia quando falliscono sia quando funzionano bene. Queste sono preoccupazioni sulla legittimità di un sistema decisionale automatizzato e le esploreremo nel Capitolo 2.

I danni dei sistemi informativi Quando un

imputato viene ingiustamente informato durante la fase istruttoria, il danno è evidente. Ma al di là del processo decisionale algoritmico, anche i sistemi informativi come gli algoritmi di ricerca e raccomandazione possono avere effetti negativi, ma qui il danno è indiretto e più difficile da definire.

Ecco un esempio. I risultati della ricerca di immagini per termini professionali come amministratore delegato o sviluppatore di software riflettono (e probabilmente esagerano) la composizione di genere prevalente e gli stereotipi su tali occupazioni.³⁷ Un altro esempio che abbiamo incontrato in precedenza sono gli stereotipi di genere nella traduzione online. Questi e altri esempi che sono inquietanti a vari livelli, come l'app di Google che etichetta le foto dei neri americani come "gorilla" o i risultati offensivi del completamento automatico, sembrano rientrare in una categoria morale diversa rispetto, ad esempio, a un sistema discriminatorio utilizzato nella giustizia penale. , che ha conseguenze immediate e tangibili.

Un discorso di Kate Crawford illustra le differenze.³⁸ Quando i sistemi decisionali nella giustizia penale, nell'assistenza sanitaria, ecc. sono discriminatori, creano danni allocativi, che vengono causati quando un sistema priva alcuni gruppi di un'opportunità o di una risorsa. Al contrario, gli altri esempi – perpetuazione degli stereotipi e denigrazione culturale – sono esempi di danni rappresentazionali, che si verificano quando i sistemi

rafforzare la subordinazione di alcuni gruppi lungo le linee dell'identità: razza, classe, genere, ecc.

I dati allocativi hanno ricevuto molta attenzione sia perché i loro effetti sono immediati, sia perché sono più facili da formalizzare e studiare in informatica ed economia. I dati rappresentazionali hanno effetti a lungo termine e resistono alla caratterizzazione formale. Ma poiché l'apprendimento automatico è diventato parte del modo in cui diamo un senso al mondo – attraverso tecnologie come la ricerca, la traduzione, gli assistenti vocali e l'etichettatura delle immagini – i dati rappresentazionali lasceranno un'impronta nella nostra cultura e influenzano la formazione dell'identità e la perpetuazione degli stereotipi. Pertanto, queste sono preoccupazioni cruciali per i campi dell'elaborazione del linguaggio naturale e della visione artificiale. Sebbene questo libro riguardi principalmente i dati allocativi, tratteremo brevemente i dati rappresentazionali nei capitoli 7 e 9.

La maggior parte dei contenuti consumati online è mediata da algoritmi di raccomandazione che influenzano quali utenti vedono quali contenuti. Pertanto, questi algoritmi influenzano quali messaggi vengono amplificati. Gli algoritmi dei social media sono stati accusati di una serie di mali: camere di risonanza in cui gli utenti sono esposti a contenuti conformi alle loro convinzioni precedenti; esacerbare la polarizzazione politica; radicalizzazione di alcuni utenti in convinzioni marginali; alimentando il risentimento etnico e la violenza; un deterioramento della salute mentale; e così via. La ricerca su queste domande è agli inizi e stabilire la causalità è difficile, e non è chiaro quanto di questi effetti siano dovuti alla progettazione dell'algoritmo rispetto al comportamento dell'utente. Ma non c'è dubbio che gli algoritmi abbiano un qualche ruolo. Twitter ha confrontato sperimentalmente un feed di contenuti non algoritmico (cronologico inverso) con un feed algoritmico e ha scoperto che i contenuti della destra politica tradizionale erano costantemente favoriti nel contesto algoritmico rispetto ai contenuti della sinistra politica tradizionale . di portata per noi. Tuttavia, parleremo brevemente della discriminazione nel targeting degli annunci pubblicitari e nei mercati online nel Capitolo 7.

La nostra prospettiva: limiti e opportunità

Abbiamo visto come l'apprendimento automatico propaghi le disuguaglianze nello stato del mondo attraverso le fasi di misurazione, apprendimento, azione e feedback. È meglio pensare ai sistemi di apprendimento automatico che influenzano le persone come a circuiti chiusi, poiché le azioni che intraprendiamo in base alle previsioni influenzano a loro volta lo stato del mondo. Uno degli obiettivi principali dell'apprendimento automatico equo è sviluppare una comprensione di quando queste disparità sono dannose, ingiustificate o altrimenti inaccettabili e sviluppare interventi per mitigare tali disparità.

Ci sono sfide e limiti fondamentali per raggiungere questo obiettivo. Una misurazione imparziale potrebbe essere irrealizzabile anche in linea di principio, come quando il costrutto stesso (ad esempio la razza) è instabile. Esistono ulteriori limitazioni pratiche derivanti dal fatto che il decisore solitamente non è coinvolto nella fase di misurazione. Inoltre, i dati osservativi possono essere insufficienti per identificare le cause delle disparità, cosa necessaria nella progettazione di interventi significativi e per comprendere gli effetti dell'intervento. La maggior parte dei tentativi di "debiasare" l'apprendimento automatico nel

la letteratura di ricerca attuale presuppone sistemi matematici semplicistici, spesso ignorando l'effetto degli interventi algoritmici sugli individui e sullo stato a lungo termine della società.

Nonostante queste importanti limitazioni, ci sono ragioni per essere cautamente ottimisti sull'equità e sull'apprendimento automatico. Innanzitutto, il processo decisionale basato sui dati ha il potenziale per essere più trasparente rispetto al processo decisionale umano. Ci costringe ad articolare i nostri obiettivi decisionali e ci consente di comprendere chiaramente i compromessi tra i desiderata. Tuttavia, ci sono sfide da superare per raggiungere questo potenziale di trasparenza. Una sfida è migliorare l'interpretabilità e la spiegabilità dei moderni metodi di apprendimento automatico, che è un argomento di vigorosa ricerca in corso. Un'altra sfida è la natura proprietaria dei set di dati e dei sistemi che sono cruciali per un dibattito pubblico informato su questo argomento.

Molti commentatori hanno invocato un cambiamento dello status quo.⁴⁰

In secondo luogo, esistono interventi efficaci in molte applicazioni di apprendimento automatico, in particolare nell'elaborazione del linguaggio naturale e nella visione artificiale. I compiti in questi ambiti (ad esempio, trascrivere il discorso) sono soggetti a un'incertezza meno intrinseca rispetto al processo decisionale tradizionale (ad esempio, prevedere se un richiedente un prestito ripagherà), eliminando alcuni dei vincoli statistici che studieremo nel Capitolo 3.

La nostra ultima e più importante ragione di ottimismo è che la svolta verso il processo decisionale automatizzato e l'apprendimento automatico offre l'opportunità di riconnettersi con i fondamenti morali dell'equità. Gli algoritmi ci costringono a essere esplicativi su ciò che vogliamo ottenere con il processo decisionale. Ed è molto più difficile nascondere le nostre intenzioni vere o poco specificate quando dobbiamo dichiarare formalmente questi obiettivi. In questo modo, l'apprendimento automatico ha il potenziale per aiutarci a discutere in modo più efficace sull'equità delle diverse politiche e procedure decisionali.

Non dovremmo aspettarci che il lavoro sull'equità nell'apprendimento automatico fornisca risposte facili. E dovremmo diffidare degli sforzi che trattano l'equità come qualcosa che può essere ridotto a un timbro di approvazione algoritmico. Dobbiamo cercare di affrontare, e non evitare, le questioni difficili quando si tratta di discutere e definire l'equità. Potrebbe anche essere necessario rivalutare il significato e l'applicabilità degli approcci esistenti alla discriminazione nella legge e nella politica,⁹ ampliando gli strumenti a nostra disposizione per ragionare sull'equità e cercare giustizia.

Ci auguriamo che questo libro possa svolgere un piccolo ruolo nello stimolare questa indagine interdisciplinare.

Note bibliografiche e approfondimenti

Questo capitolo attinge da diverse tassonomie di pregiudizi nell'apprendimento automatico e nel processo decisionale basato sui dati: un post sul blog di Moritz Hardt,²⁵ un articolo di Barocas e Selbst,⁹ e un rapporto del 2016 dell'Office of Science and Technology Policy della Casa Bianca.⁴¹ Per un'ampia rassegna delle sfide sollevate dall'intelligenza artificiale, dall'apprendimento automatico e dai sistemi algoritmici, vedere il rapporto AI Now.⁴² Uno dei primi

lavori che ha studiato l'equità nei sistemi algoritmici è quello di Friedman e Nissenbaum nel 1996.⁴³ Articoli che studiano le disparità demografiche nella classificazione

tion ha iniziato ad apparire regolarmente a partire dal 2008;⁴⁴ il luogo di questa ricerca era in Europa e nella comunità di ricerca sul data mining. Con l'istituzione del workshop FAT/ML nel 2014, è emersa una nuova comunità e da allora l'argomento è diventato sempre più popolare. Diversi libri di pubblico popolare hanno fornito critiche ai sistemi algoritmici nella società moderna: *The Black Box Society* di Frank Pasquale,⁴⁵ *Weapons of Math Destruction* di Cathy O'Neill,⁴⁶ *Automating inequality* di Virginia Eubanks,⁴⁷ e *Algorithms of Oppression* di Safiya Noble.⁴⁸

2

Quando il processo decisionale automatizzato è legittimo?

Questi tre scenari hanno qualcosa in comune:

- Una studentessa è orgogliosa del saggio creativo che ha scritto per un test standardizzato. Riceve un punteggio perfetto, ma è delusa nell'apprendere che il test è stato effettivamente valutato da un computer.
- Un imputato ritiene che un sistema di previsione del rischio penale lo abbia classificato come ad alto rischio di mancata comparizione in tribunale, sulla base del comportamento di altri come lui, nonostante avesse tutta l'intenzione di comparire in tribunale alla data stabilita.
- Un sistema automatizzato ha bloccato un utente di un social media per aver violato la politica della piattaforma sul comportamento accettabile. L'utente insiste di non aver fatto nulla di male, ma la piattaforma non fornirà ulteriori dettagli né alcuna procedura di ricorso.

Tutti questi sono processi decisionali automatizzati o sistemi di supporto decisionale che probabilmente sembrano ingiusti o ingiusti. Eppure si tratta di un senso di ingiustizia diverso da quello di cui abbiamo parlato nel primo capitolo (e su cui torneremo nel prossimo capitolo). Non si tratta del trattamento relativo dei diversi gruppi. Ciò che riguarda invece queste domande è la legittimità: se sia giusto implementare un sistema del genere in un dato scenario. Questa domanda, a sua volta, influenza sulla legittimità dell'organizzazione che lo implementa.

La maggior parte delle istituzioni ha bisogno di legittimità per poter funzionare in modo efficace. Le persone devono credere che l'istituzione sia ampiamente allineata ai valori sociali. La ragione di ciò è relativamente chiara nel caso di istituzioni pubbliche come il governo o le scuole, che sono direttamente o indirettamente responsabili nei confronti del pubblico.

È meno chiaro il motivo per cui le imprese private abbiano bisogno di legittimità. Una risposta è che quanto maggiore è il potere che un'impresa ha sugli individui, tanto più l'esercizio di tale potere deve essere percepito come legittimo. E il processo decisionale sulle persone implica l'esercizio del potere su di loro, quindi è importante garantirne la legittimità. Altrimenti, le persone troveranno vari modi per resistere, in particolare attraverso la legge. Una perdita di legittimità potrebbe anche compromettere la capacità di un'impresa di competere sul mercato.

Nel settore della tecnologia digitale sono emerse ripetutamente domande sulla legittimità delle imprese. Ad esempio, le società di ride sharing hanno dovuto affrontare tali questioni, portando ad attivismo, contenziosi e regolamentazione. Imprese i cui modelli di business si basano

Anche i dati personali, soprattutto quelli raccolti di nascosto, hanno subito crisi di percezione. Oltre alle risposte legali, tali aziende hanno visto i concorrenti trarre vantaggio dalle loro pratiche permissive in materia di privacy. Ad esempio, Apple ha reso più difficile per Facebook monitorare gli utenti su iOS, incidendo negativamente sulle sue entrate.⁴⁹ Questa mossa ha goduto del sostegno pubblico nonostante le vociferanti proteste di Facebook, probabilmente perché il modello di business sottostante aveva perso legittimità.

Per queste ragioni, un libro sull'equità è incompleto senza una discussione sulla legittimità. Inoltre, la questione della legittimità dovrebbe precedere le altre questioni di equità. Discutere sulla giustizia distributiva nel contesto di un'istituzione fondamentalmente ingiusta è, nella migliore delle ipotesi, una perdita di tempo, e nel peggiore dei casi aiuta a sostenere la legittimità dell'istituzione, ed è quindi controproducente. Ad esempio, il miglioramento della tecnologia di analisi facciale per ridurre la disparità nei tassi di errore tra i gruppi razziali non è una risposta utile alle preoccupazioni sull'uso di tali tecnologie per scopi oppressivi.⁵⁰

Le discussioni sulla legittimità sono state in gran parte oscurate dalle discussioni sui pregiudizi e sulla discriminazione nel discorso sull'equità. Spesso i sostenitori hanno scelto di concentrarsi su considerazioni distributive come un modo per attaccare la legittimità, poiché tende ad essere un'argomentazione più semplice da sostenere. Ma questo può rivelarsi controproducente, poiché molte aziende hanno cooptato il discorso sull'equità e trovano relativamente facile garantire la parità nelle decisioni tra gruppi demografici senza affrontare le preoccupazioni sulla legittimità.⁵¹

Questo capitolo riguarda la legittimità: se è moralmente giustificabile l'uso apprendimento automatico o metodi automatizzati in un dato scenario.

Sebbene abbiano sottolineato l'importanza fondamentale della legittimità, i lettori interessati alle questioni distributive possono saltare al capitolo 3 per una trattazione tecnica o al capitolo 4 per una spiegazione normativa; quei capitoli, il capitolo 3 in particolare, non si basano direttamente su questo.

L'apprendimento automatico non sostituisce il processo decisionale umano

L'apprendimento automatico svolge un ruolo importante nelle decisioni che allocano risorse e opportunità fondamentali per le possibilità di vita delle persone. La posta in gioco è chiaramente alta. Ma è da molto tempo che le persone prendono decisioni ad alto rischio l'una rispetto all'altra, e tali decisioni sembrano essere soggette a un esame molto meno critico.

Ecco una visione di paglia: le decisioni basate sull'apprendimento automatico sono analoghe al processo decisionale degli esseri umani, e quindi l'apprendimento automatico non merita particolare attenzione. Anche se è vero che i modelli di apprendimento automatico potrebbero essere difficili da comprendere per le persone, anche gli esseri umani sono scatole nere. E sebbene possano esserci pregiudizi sistematici nei modelli di apprendimento automatico, spesso sono evidentemente meno distorti degli esseri umani.

Rifiutiamo questa analogia tra l'apprendimento automatico e il processo decisionale umano. Comprendendo perché fallisce e quali analogie sono più appropriate, svilupperemo una migliore comprensione di ciò che rende l'apprendimento automatico particolarmente pericoloso come modo per prendere decisioni ad alto rischio.

Sebbene l'apprendimento automatico venga talvolta utilizzato per automatizzare i compiti eseguiti nella testa di un essere umano, molte delle decisioni ad alto rischio su cui si concentrano

il lavoro sull'equità e sull'apprendimento automatico è quello tradizionalmente svolto dalle burocrazie. Ad esempio, le decisioni relative ad assunzioni, crediti e ammissioni sono raramente lasciate a una persona che può prenderle da sola come ritiene opportuno. Invece, queste decisioni sono guidate da regole e procedure formali, che coinvolgono molti attori con ruoli e responsabilità prescritti. La burocrazia è nata in parte come risposta alla soggettività, all'arbitrarietà e all'incoerenza del processo decisionale umano; le sue regole e procedure istituzionalizzate mirano a minimizzare gli effetti delle fragilità degli esseri umani come decisori individuali.⁵²

Naturalmente, le burocrazie non sono perfette. Il termine stesso burocrazia tende ad avere una connotazione negativa: un processo inutilmente contorto che è difficile o impossibile da gestire. E nonostante il loro approccio eccessivamente formalistico (si potrebbe dire freddo) al processo decisionale, le burocrazie raramente riescono a disciplinare pienamente i singoli decisori che occupano le loro fila. Le burocrazie rischiano di essere altrettanto capricciose e imperscrutabili degli esseri umani, ma molto più disumanizzanti.⁵²

Ecco perché le burocrazie spesso incorporano tutele procedurali: meccanismi che garantiscono che le decisioni siano prese in modo trasparente, sulla base di informazioni giuste e pertinenti, e con l'opportunità di contestazioni e correzioni. Una volta che ci rendiamo conto che l'apprendimento automatico viene utilizzato per automatizzare le decisioni burocratiche piuttosto che individuali, affermare che gli esseri umani non hanno bisogno – o semplicemente non possono – rendere conto delle loro decisioni quotidiane non esenta l'apprendimento automatico da queste aspettative. Come ha sostenuto Katherine Strandburg, “[la]ragionamento è un requisito fondamentale nei sistemi decisionali convenzionali proprio perché i decisori umani sono imperscrutabili e inclini a pregiudizi ed errori, non a causa di alcuna aspettativa che essi forniranno, o addirittura potranno, fornire informazioni accurate ed affidabili. descrizioni

dettagliate dei loro processi mentali”.⁵³ Analizzando l'apprendimento automatico al processo decisionale burocratico – piuttosto che individuale – possiamo comprendere meglio l'origine di alcune delle preoccupazioni sull'apprendimento automatico. Quando viene utilizzato in ambiti ad alto rischio, mina il tipo di protezione che spesso mettiamo in atto per garantire che le burocrazie siano coinvolte in un processo decisionale ben eseguito e ben giustificato.

La burocrazia come baluardo contro il processo decisionale arbitrario

Il tipo di processo decisionale problematico da cui le burocrazie proteggono può essere definito processo decisionale arbitrario. Kathleen Creel e Deborah Hellman hanno utilmente distinto tra due tipi di arbitrarietà.⁵⁴ In primo luogo, l'arbitrarietà potrebbe riferirsi a decisioni prese su una base incoerente o ad hoc. In secondo luogo, l'arbitrarietà potrebbe riferirsi alla base per il processo decisionale priva di motivazione, anche se le decisioni vengono prese coerentemente su tale base. Questa prima visione dell'arbitrarietà riguarda principalmente la regolarità procedurale⁵⁵: se uno schema decisionale viene eseguito in modo coerente e corretto. Le preoccupazioni sull'arbitrarietà, in questo caso, riguardano in realtà se le regole che governano le decisioni importanti siano fissate in anticipo e applicate in modo appropriato, con l'obiettivo di ridurre la capacità dei decisorи di prendere decisioni in modo casuale.

Quando il processo decisionale è arbitrario in questo senso del termine, gli individui possono scoprire di essere soggetti a diversi schemi decisionali e di ricevere decisioni diverse semplicemente perché capita di attraversare il processo decisionale in momenti diversi. Non solo lo schema decisionale potrebbe cambiare nel tempo; i decisori umani potrebbero essere incoerenti nel modo in cui applicano questi schemi mentre affrontano casi diversi. Quest'ultimo potrebbe essere vero per un singolo decisore il cui comportamento è incoerente nel tempo, ma potrebbe anche essere vero se il processo decisionale assegna i casi a diversi individui che sono individualmente coerenti, ma differiscono tra loro. Pertanto, anche due persone identiche per quanto riguarda i criteri decisionali possono ricevere decisioni diverse, violando l'aspettativa che persone simili debbano essere trattate in modo simile quando si tratta di decisioni ad alto rischio.

Questo principio si basa sulla convinzione che le persone hanno diritto a decisioni simili a meno che non ci siano ragioni per trattarle in modo diverso (affronteremo presto ciò che determina se queste sono buone ragioni). Per decisioni particolarmente consequenziali, le persone possono avere buone ragioni per chiedersi perché qualcuno che gli somigliava ha ricevuto il risultato desiderato dal processo decisionale mentre loro no.

L'incoerenza è problematica anche quando impedisce alle persone di sviluppare piani di vita efficaci basati sulle aspettative sui sistemi decisionali che devono navigare per ottenere risorse e opportunità desiderabili.⁵⁴ Pertanto, un processo decisionale incoerente è ingiusto sia perché potrebbe comportare differenze ingiustificate trattamento di individui simili e anche perché rappresenta una minaccia all'autonomia individuale impedendo alle persone di prendere decisioni efficaci su come perseguire al meglio i propri obiettivi di vita.

La seconda visione dell'arbitrarietà tocca una questione più profonda: ci sono buone ragioni – o qualche ragione – per cui lo schema decisionale appare così com'è? Ad esempio, se un allenatore sceglie una squadra di atletica in base al colore delle scarpe da ginnastica dei corridori, ma lo fa in modo coerente, è comunque arbitrario perché il criterio non ha una base valida. Non aiuta a portare avanti gli obiettivi del decisore (ad esempio, mettere insieme una squadra di corridori che vincerà l'incontro imminente).

L'arbitrarietà, da questa prospettiva, è problematica perché mina una giustificazione fondamentale per lo schema decisionale scelto: il fatto che esso effettivamente aiuti a portare avanti gli obiettivi del decisore. Se lo schema decisionale non fa nulla per raggiungere questi obiettivi, allora non c'è alcuna ragione giustificata per optare per quello schema decisionale e per trattare le persone di conseguenza. Quando le risorse e le opportunità desiderabili vengono allocate arbitrariamente, si sottopongono inutilmente gli individui a decisioni diverse, nonostante il fatto che tutti gli individui possano avere uguale interesse in queste risorse e opportunità.

Nel contesto del processo decisionale del governo, c'è spesso un requisito legale che ci sia una base razionale per il processo decisionale, cioè che ci siano buone ragioni per prendere le decisioni nel modo in cui sono.⁵⁴ Regole che non aiutano il governo raggiungere gli obiettivi politici dichiarati entra in conflitto con i principi del giusto processo. Ciò potrebbe essere dovuto al fatto che le regole sono state scelte arbitrariamente o a causa di qualche evidente difetto nel ragionamento alla base di queste regole. Questi requisiti derivano dal fatto che il governo ha il monopolio su alcuni aspetti altamente consequenziali

decisioni, lasciando le persone senza alcuna possibilità di ricorrere in giudizio affrontando il proprio caso con un altro decisore.

Non esiste alcun obbligo legale corrispondente quando i decisori sono attori privati, come sottolineano Creel e Hellman. Le aziende sono spesso libere di prendere decisioni scarsamente ragionate, persino del tutto arbitrarie. In teoria, gli schemi decisionali che sembrano non fare nulla per promuovere gli obiettivi degli attori privati dovrebbero essere espulsi dal mercato da schemi concorrenti che sono più efficaci.⁵⁴

Nonostante ciò, spesso ci aspettiamo che le decisioni di maggiore importanza, anche quando prese da attori privati, siano prese per buone ragioni. Non tollereremo probabilmente datori di lavoro, istituti di credito o funzionari di ammissione che prendono decisioni sui candidati lanciando una moneta o in base al colore delle scarpe da ginnastica dei candidati. Perché potrebbe essere questo?

Il processo decisionale arbitrario non rispetta la gravità di queste decisioni e mostra una mancanza di rispetto per le persone ad esse soggette. Anche se accettiamo di non poter dettare gli obiettivi delle istituzioni, siamo comunque contrari a dare loro la completa libertà di trattare le persone come preferiscono. Quando la posta in gioco è sufficientemente alta, i decisori si assumono l'onere di giustificare i loro schemi decisionali nel rispetto degli interessi delle persone interessate da tali decisioni. Il fatto che le persone possano tentare la fortuna con altri decisori nello stesso ambito (ad esempio, un altro datore di lavoro, un finanziatore o un funzionario di ammissione) può fare ben poco per modulare queste aspettative.

Tre forme di automazione

Per ricapitolare la nostra discussione precedente, l'automazione potrebbe minare importanti tutele procedurali nel processo decisionale burocratico. Ma cosa aiuta esattamente il machine learning ad automatizzare? Si scopre che esistono tre diversi tipi di automazione.

Il primo tipo di automazione implica l'adozione di regole decisionali stabilite manualmente (ad esempio, elaborate attraverso un processo di policy-making tradizionale) e la loro traduzione in software, con l'obiettivo di automatizzare la loro applicazione a casi particolari.⁵⁵ Ad esempio, molte agenzie governative seguono questo approccio quando adottano software per automatizzare le determinazioni sull'ammissibilità dei benefici in conformità con le politiche preesistenti. Allo stesso modo, i datori di lavoro seguono questo approccio quando identificano determinate qualifiche minime per un lavoro e sviluppano software per rifiutare automaticamente i candidati che non le possiedono. In entrambi i casi le regole sono ancora stabilite dall'uomo, ma la loro applicazione è automatizzata da un computer; l'apprendimento automatico non ha un ruolo ovvio qui.

Ma che dire dei casi in cui i decisori umani si sono affidati principalmente al giudizio informale piuttosto che a regole formalmente specificate? È qui che entra in gioco il secondo tipo di automazione. Utilizza l'apprendimento automatico per capire come replicare i giudizi informali degli esseri umani. Avendo scoperto automaticamente uno schema decisionale che produce le stesse decisioni prese dagli esseri umani in passato, implementa quindi questo schema nel software per sostituire gli umani che

aveva preso queste decisioni. Lo studente il cui saggio creativo è stato sottoposto a valutazione computerizzata, descritto all'inizio di questo capitolo, è un esempio di tale approccio: il software in questo caso cerca di replicare le valutazioni soggettive dei valutatori umani.

Il tipo finale di automazione è abbastanza diverso dai primi due. Non si basa su uno schema decisionale burocratico esistente o sul giudizio umano. Si tratta invece di apprendere le regole decisionali dai dati. Utilizza un computer per scoprire modelli in un set di dati che prevedono un risultato o una proprietà di interesse politico, quindi basa le decisioni su tali previsioni. Si noti che tali regole potrebbero essere applicate manualmente (dagli esseri umani) o automaticamente (tramite software). Il punto rilevante dell'automazione, in questo caso, è nel processo di sviluppo delle regole, non necessariamente nella loro applicazione. Ad esempio, queste potrebbero essere regole che impongono alla polizia di pattugliare determinate aree, date le previsioni sulla probabile incidenza della criminalità basate su osservazioni passate di crimini. Oppure potrebbero essere regole che suggeriscono che gli istituti di credito concedano credito a determinati richiedenti, date le storie di rimborso di precedenti beneficiari come loro. L'apprendimento automatico – e altre tecniche statistiche – sono cruciali per questa forma di automazione.

Come vedremo nelle prossime tre sezioni, ogni tipo di automazione solleva preoccupazioni specifiche.

Automatizzare le regole decisionali preesistenti

Per molti aspetti, la prima forma di automazione – tradurre regole preesistenti in software in modo che le decisioni possano essere eseguite automaticamente – è una risposta diretta all'arbitrarietà intesa come incoerenza. L'automazione aiuta a garantire la coerenza nel processo decisionale perché richiede che lo schema per prendere le decisioni sia fissato. Significa anche che lo schema viene applicato ogni volta nello stesso modo.

Eppure molte cose possono andare storte. Danielle Citron offre un resoconto convincente dei pericoli derivanti dall'automazione delle regole decisionali stabilite attraverso un processo di elaborazione politica o di regolamentazione deliberativa.⁵⁶ Automatizzare l'esecuzione di uno schema decisionale preesistente richiede la traduzione di tale schema in codice. I programmatore potrebbero commettere errori in tale processo, portando a decisioni automatizzate che divergono dalla politica che il software dovrebbe eseguire. Un altro problema è che la politica che i programmatore hanno il compito di automatizzare potrebbe non essere sufficientemente esplicita o precisa; di fronte a tale ambiguità, i programmatore potrebbero assumersi la responsabilità di esprimere il proprio giudizio, usurpando di fatto l'autorità di definire la politica. E al livello più elementare, il software potrebbe essere difettoso. Ad esempio, centinaia di direttori delle poste britannici sono stati condannati per furto o frode nell'arco di vent'anni sulla base di software difettoso, in quello che è stato definito il più grande errore giudiziario nella storia britannica.⁵⁷

Automatizzare il processo decisionale può anche essere problematico quando elimina completamente ogni spazio di discrezionalità. Sebbene la discrezionalità umana presenti i propri problemi, come descritto sopra, può essere utile quando è difficile o impossibile specificare completamente come dovrebbero essere prese le decisioni in conformità con gli obiettivi e i principi dell'istituzione.⁵⁸ L'automazione richiede che un'istituzione determini in anticipo tutto di

i criteri di cui uno schema decisionale terrà conto; non c'è spazio per considerare la rilevanza di dettagli aggiuntivi che potrebbero non essere stati considerati o previsti al momento dello sviluppo del software.

È quindi probabile che il processo decisionale automatizzato sia molto più fragile del processo decisionale che implica una revisione manuale perché limita l'opportunità per i soggetti decisionali di introdurre informazioni nel processo decisionale. Le persone si limitano a fornire prove che corrispondano a un campo prestabilito nel software. Tali vincoli possono portare a situazioni assurde in cui la rigorosa applicazione delle regole decisionali porta a risultati direttamente contrari agli obiettivi dietro queste regole. Nuove prove che invertirebbero immediatamente la valutazione di un decisore umano potrebbero non trovare posto nel processo decisionale automatizzato.⁵⁹ Ad esempio, in un sistema automatizzato per valutare le persone malate per determinare l'idoneità a un caregiver fornito dallo Stato, un campo chiedeva se esiste ci fossero problemi ai piedi. Un perito ha visitato una certa persona e ha compilato il campo per indicare che non aveva alcun problema, perché era amputata.⁶⁰

La discrezione è preziosa in questi casi perché gli esseri umani sono spesso in grado di riflettere sulla rilevanza delle informazioni aggiuntive per la decisione in questione e sull'obiettivo sottostante a cui tali decisioni dovrebbero servire. In effetti, la revisione umana lascia spazio per espandere i criteri in esame e per riflettere su quando l'applicazione meccanica delle regole non riesce a raggiungere lo scopo previsto.^{61, 59}

Questi stessi vincoli possono anche limitare la capacità delle persone di segnalare errori o di contestare la decisione finale.⁶² Quando interagisce con un funzionario addetto ai prestiti, una persona potrebbe far notare che la sua pratica di credito contiene informazioni errate. Quando richiedono un prestito tramite un processo automatizzato, potrebbero non avere opportunità equivalenti. O forse una persona riconosce che le regole che stabiliscono la sua ammissibilità ai benefici governativi sono state applicate in modo errato. Quando gli operatori sociali vengono sostituiti dal software, le persone soggette a queste decisioni potrebbero non avere i

mezzi per sollevare obiezioni giustificate.⁶³ Infine, l'automazione corre il serio rischio di limitare la responsabilità e di esacerbare gli effetti disumanizzanti del trattare con le burocrazie. L'automazione può rendere difficile identificare l'agente responsabile di una decisione; il software spesso ha l'effetto di disperdere il luogo della responsabilità perché la decisione sembra non essere presa da nessuno.⁶⁴ Le persone possono disporre di mezzi più efficaci per contestare le decisioni e contestare lo schema decisionale quando il processo decisionale è affidato a persone identificabili. Allo stesso modo, la capacità dell'automazione di rimuovere gli esseri umani dal processo decisionale può contribuire a far sì che le persone abbiano la sensazione che un'istituzione non le consideri degne del rispetto che garantirebbe loro l'opportunità di apportare correzioni legittime, introdurre ulteriori informazioni rilevanti o descrivere circostanze attenuanti. .⁶⁵ Questo è proprio il problema evidenziato dall'esempio iniziale di un utente di un social media che è stato espulso da una piattaforma senza spiegazioni o possibilità di appello.

Abbiamo evidenziato molte preoccupazioni normative che derivano dalla semplice automazione dell'applicazione di uno schema decisionale preesistente. Sebbene molti di questi problemi siano comunemente attribuiti all'adozione dell'apprendimento automatico, nessuno di essi ha origine dall'uso specifico dell'apprendimento automatico. Sforzi di lunga data per

automatizzare il processo decisionale con il software tradizionale comporta di per sé molti pericoli. Il fatto che l'apprendimento automatico non sia la causa esclusiva di questo tipo di problemi non è un motivo per prenderli meno sul serio, ma risposte efficaci a questi problemi richiedono che siamo chiari sulle loro origini.

Apprendere le regole decisionali dai dati sulle decisioni passate per automatizzare Loro

I decisori potrebbero avere un processo preesistente ma informale per prendere decisioni che potrebbero voler automatizzare. In questo caso, l'apprendimento automatico (o altre tecniche statistiche) potrebbe essere impiegato per "prevedere" come un essere umano prenderebbe una decisione, dati determinati criteri. L'obiettivo non è necessariamente quello di recuperare perfettamente il peso specifico che i decisori del passato avevano implicitamente assegnato a criteri diversi, ma piuttosto quello di garantire che il modello produca un insieme di decisioni simili a quelle umane. Per tornare a uno dei nostri esempi ricorrenti, un istituto scolastico potrebbe voler automatizzare il processo di valutazione dei saggi, e potrebbe tentare di farlo affidandosi all'apprendimento automatico per imparare a imitare i voti che gli insegnanti hanno assegnato a lavori simili in passato.

Questa forma di automazione potrebbe aiutare ad affrontare le preoccupazioni relative all'arbitrarietà nel processo decisionale umano formalizzando e fissando uno schema decisionale simile a quello che gli esseri umani avrebbero potuto impiegare in passato. A questo riguardo, l'apprendimento automatico potrebbe essere auspicabile perché può aiutare a appianare eventuali incoerenze nelle decisioni umane da cui ha indotto alcune regole decisionali. Ad esempio, il modello di valutazione del saggio sopra descritto potrebbe ridurre parte della varianza osservata nella valutazione degli insegnanti le cui valutazioni soggettive il modello sta imparando a replicare. L'automazione può ancora una volta aiutare ad affrontare le preoccupazioni relative all'arbitrarietà intesa come incoerenza, anche quando sono i giudizi soggettivi ad essere automatizzati.

Qualche decennio fa, esisteva un approccio popolare all'automazione che si basava sulla codifica esplicita del ragionamento su cui gli esseri umani facevano affidamento per prendere decisioni.⁶⁶ Questo approccio, chiamato sistemi esperti, fallì per molte ragioni, compreso il fatto che le persone non sono sempre in grado di spiegare il proprio ragionamento.⁶⁷ I sistemi esperti alla fine cedettero il passo all'approccio che consisteva semplicemente nel chiedere alle persone di etichettare gli esempi e nel far sì che algoritmi di apprendimento scoprissero come prevedere al meglio l'etichetta che gli umani avrebbero assegnato. Anche se questo approccio si è rivelato efficace, presenta i suoi pericoli.

In primo luogo, potrebbe dare una patina di valutazione oggettiva a schemi decisionali che semplicemente automatizzano il giudizio soggettivo degli esseri umani. Di conseguenza, le persone potrebbero essere più propense a considerare le sue decisioni come meno meritevoli di indagine critica. Ciò è particolarmente preoccupante perché apprendere le regole decisionali dalle precedenti decisioni prese dagli esseri umani corre l'ovvio rischio di replicare ed esagerare qualsiasi qualità discutibile del processo decisionale umano imparando dai cattivi esempi forniti dagli esseri umani. (In effetti, molti tentativi di apprendere una regola per prevedere qualche obiettivo di interesse apparentemente oggettivo – la forma di automazione di cui parleremo nella prossima sezione – sono in realtà solo una versione di replica del giudizio umano sotto mentite spoglie. Se non possiamo ottenere una verità oggettiva per il target scelto

previsione, non c'è modo di sfuggire al giudizio umano. Come sottolinea David Hand, gli esseri umani spesso dovranno esercitare discrezione nello specificare e identificare ciò che conta come esempio dell'obiettivo.⁶⁸⁾

In secondo luogo, tali schemi decisionali possono essere considerati equivalenti a quelli impiegati dagli esseri umani e quindi suscettibili di funzionare allo stesso modo, anche se il modello potrebbe raggiungere le sue decisioni in modo diverso e produrre modelli di errore del tutto diversi.⁶⁹⁾ Anche quando il modello è in grado di prevedere con un alto grado di precisione le decisioni che gli esseri umani prenderebbero dato un particolare input, non vi è alcuna garanzia che il modello abbia ereditato tutte le sfumature e le considerazioni che entrano nel processo decisionale umano. Peggio ancora, i modelli potrebbero anche imparare a fare affidamento su criteri in modi che gli esseri umani troverebbero preoccupanti o discutibili, anche se così facendo si produce comunque un insieme di decisioni simili a quelle che gli esseri umani prenderebbero.⁷⁰⁾ Ad esempio, un modello che automatizza la valutazione del saggio assegnando punteggi più alti agli elaborati che utilizzano un vocabolario sofisticato possono fare un lavoro ragionevolmente buono replicando i giudizi dei valutatori umani (probabilmente perché una scrittura di qualità superiore tende a fare affidamento su un vocabolario più sofisticato), ma è improbabile che controllare la presenza di determinate parole sia un sostituto affidabile per valutare un saggio di coerenza logica e correttezza fattuale.⁷¹⁾

In breve, l'uso dell'apprendimento automatico per automatizzare le decisioni precedentemente prese dagli esseri umani può essere problematico perché può finire per essere troppo simile ai decisori umani e troppo diverso da loro.

Derivare regole decisionali imparando a prevedere un obiettivo

L'ultima forma di automazione è quella in cui i decisori si affidano all'apprendimento automatico per apprendere una regola o una politica decisionale dai dati. Questa forma di automazione, che chiameremo ottimizzazione predittiva, risponde direttamente alle preoccupazioni relative al processo decisionale ragionato. Si noti che nessuna delle prime due forme di automazione lo fa. L'esecuzione coerente di una politica preesistente tramite l'automazione non garantisce che la politica stessa sia ragionata. Né fare affidamento sulle decisioni umane passate per indurre una regola decisionale garantisce che la base per il processo decisionale automatizzato rifletterà giudizi ragionati. In entrambi i casi, lo schema decisionale sarà ragionato solo quanto la politica formale o i giudizi informali la cui esecuzione è automatizzata.

Al contrario, l'ottimizzazione predittiva cerca di fornire una base più rigorosa per il processo decisionale basandosi solo su criteri nella misura in cui essi predicono in modo dimostrabile il risultato o la qualità dell'interesse. Se impiegato in questo modo, l'apprendimento automatico sembra garantire decisioni ragionate perché i criteri che sono stati incorporati nello schema decisionale – e la loro particolare ponderazione – sono dettati da quanto bene prevedono l'obiettivo. E finché l'obiettivo scelto è un buon indicatore degli obiettivi dei decisori, fare affidamento su criteri che predicono questo obiettivo per prendere decisioni sembrerebbe ben motivato perché ciò aiuterà a raggiungere gli obiettivi dei decisori.

A differenza delle prime due forme di automazione, l'ottimizzazione predittiva rappresenta un allontanamento radicale dall'approccio tradizionale al processo decisionale. Nel tradizionale

approccio, un insieme di decisori ha un obiettivo – anche se questo obiettivo è amorpho e difficile da specificare – e vorrebbe sviluppare uno schema decisionale esplicito per contribuire a realizzare il proprio obiettivo. Si impegnano in discussioni e deliberazioni per cercare di raggiungere un accordo sui criteri rilevanti per la decisione e sul peso da assegnare a ciascun criterio nello schema decisionale. Facendo affidamento sull'intuizione, sulle prove pregresse e sul ragionamento normativo, i decisori sceglieranno e combineranno le caratteristiche in modi che si pensa possano aiutare a realizzare i loro obiettivi.

L'approccio statistico o di machine learning funziona diversamente. In primo luogo, i decisori cercano di identificare un obiettivo esplicito per la previsione che considerano sinonimo del loro obiettivo – o un ragionevole proxy per esso. In uno scenario di ammissione al college, un obiettivo potrebbe essere il rendimento scolastico al college e il GPA del college potrebbe esserne un indicatore. Una volta risolto questo problema, i decisori utilizzano i dati per scoprire quali criteri utilizzare e come ponderarli per prevedere al meglio l'obiettivo. Anche se potrebbero esercitare discrezionalità nella scelta dei criteri da utilizzare, la ponderazione di questi criteri sarebbe dettata interamente dall'obiettivo di massimizzare l'accuratezza della previsione risultante dell'obiettivo scelto. In altre parole, le regole decisionali verrebbero, in gran parte, apprese dai dati, piuttosto che stabilite in base alle intuizioni soggettive, alle aspettative e agli impegni normativi dei decisori.

Tabella 2.1: Confronto tra il processo decisionale tradizionale e l'ottimizzazione predittiva

	Approccio tradizionale	Approccio di ottimizzazione predittiva
Esempio: ammissione all'università	Approccio olistico che tiene conto dei risultati ottenuti, del carattere, delle circostanze speciali e di altri fattori	Addestrare un modello basato sui dati degli studenti precedenti per prevedere il GPA dei candidati se ammessi; ammettere i candidati con il punteggio più alto
Obiettivo e target	Nessun obiettivo esplicito; l'obiettivo è implicito (e di solito ci sono più obiettivi)	Definire un target esplicito; supponiamo che sia un buon proxy per l'obiettivo
Focus della deliberazione	Il dibattito riguarda il modo in cui i criteri dovrebbero influenzare la decisione	Il dibattito riguarda in gran parte la scelta dell'obiettivo
Efficacia	Potrebbe non riuscire a produrre regole che soddisfino i loro presunti obiettivi	L'accuratezza predittiva può essere quantificata
Gamma di normativo considerazioni	È più facile incorporare più principi normativi come la necessità	Più difficile incorporare più principi normativi
Giustificazione	Può essere difficile indovinare le ragioni che spingono i legislatori a scegliere un determinato schema decisionale	Le ragioni dello schema decisionale scelto sono rese esplicite nella scelta dell'obiettivo

Ciascun approccio presenta vantaggi e svantaggi dal punto di vista normativo. L'approccio tradizionale consente di esprimere molteplici obiettivi e valori normativi attraverso la scelta dei criteri e il peso ad essi assegnato.

Nell'approccio del machine learning, nella scelta del target è necessario includere molteplici obiettivi e considerazioni normative. Nelle ammissioni al college, tali obiettivi e considerazioni potrebbero includere, oltre al potenziale scolastico, il potenziale atletico e di leadership, la misura in cui il candidato contribuirebbe alla vita del campus, se il richiedente porta esperienze di vita insolite, il suo grado di bisogno e molti altri. . L'approccio più comune è quello di definire una variabile target composita che combini linearmente più componenti, ma questo diventa rapidamente ingombrante ed è raramente soggetto a un dibattito approfondito. C'è anche un certo margine per esercitare un giudizio normativo sulla scelta di includere o escludere determinati criteri decisionali, ma siamo ben lontani dal processo decisionale di politica deliberativa.

D'altra parte, se crediamo che un obiettivo, in effetti, copra l'intera gamma di obiettivi che i decisori hanno in mente, i modelli di machine learning potrebbero essere in grado di raggiungere questi obiettivi in modo più efficace. Ad esempio, in un articolo che mette a confronto i due approcci al policy making, Rebecca Johnson e Simone Zhang mostrano che l'approccio tradizionale (cioè la creazione manuale di regole attraverso un processo di deliberazione e dibattito) spesso non riesce a produrre regole che soddisfino i loro presunti obiettivi.⁷² Nell'esaminare le regole per l'assegnazione dell'assistenza abitativa, si scopre che le autorità per l'edilizia abitativa danno priorità ai veterani rispetto alle famiglie particolarmente gravate dall'affitto , nonostante il fatto che il sostegno a tali famiglie sembra essere più in linea con l'obiettivo fondamentale della politica. Johnson e Zhang affermano che mentre questa definizione delle priorità potrebbe essere l'intento reale dei politici che stabiliscono le regole, le ragioni di questa priorità sono raramente rese esplicite nel processo di deliberazione e sono particolarmente difficili da discernere dopo il fatto. Se queste regole fossero sviluppate invece utilizzando l'apprendimento automatico, i politici dovrebbero concordare un obiettivo esplicito di previsione, il che lascerebbe molto meno spazio alla confusione sulle intenzioni dei politici. E garantirebbe che le regole risultanti siano progettate solo per prevedere quell'obiettivo.⁷² Come hanno sostenuto Rediet Abebe, Solon Barocas, Jon Kleinberg e colleghi, “[I]lja natura dell'informatica è tale da richiedere scelte esplicite sugli input, obiettivi, vincoli e presupposti di un sistema”⁷³ – e questo può essere positivo se porta allo scoperto alcune considerazioni politiche e giudizi normativi.

L'approccio del machine learning corre tuttavia il serio rischio di concentrarsi esclusivamente sull'accuratezza delle previsioni. In altre parole, le decisioni “buone” sono quelle che prevedono accuratamente l'obiettivo. Ma il processo decisionale potrebbe essere “buono” per altri motivi: concentrarsi sulle qualità o sui risultati giusti (in altre parole, l' obiettivo è un buon indicatore dell'obiettivo), considerare solo i fattori rilevanti, considerare l'insieme completo dei fattori rilevanti, incorporare altri principi normativi (ad esempio, bisogno, deserto, ecc.), o consentire alle persone di comprendere e potenzialmente contestare la politica.

Anche un processo decisionale che non sia particolarmente accurato potrebbe essere considerato positivo se possiede alcune di queste altre proprietà.⁷⁴ Nelle prossime sezioni esploreremo come ciascuna di queste preoccupazioni potrebbe applicarsi all'apprendimento automatico.

Discrepanza tra target e obiettivo

Identificare un obiettivo di previsione che corrisponda bene agli obiettivi del decisore è raramente semplice. I decisori spesso non hanno in mente un obiettivo preesistente, chiaro e discreto.⁷⁵ Quando lo fanno, l'obiettivo può essere molto più complesso e sfaccettato di un risultato discreto e facilmente misurabile.⁷⁶ In effetti, i decisori possono avere molteplici obiettivi contrastanti, che forse comportano alcuni compromessi tra di loro. Ad esempio, gli schemi decisionali adottati dai funzionari addetti all'ammissione all'università spesso codificano una serie di obiettivi diversi. Non si limitano a classificare i candidati in base alla media dei voti prevista e poi ammettono i migliori candidati a riempire la capacità fisica della scuola. A parte il fatto ovvio che ciò favorirebbe i candidati che seguono corsi "facili", gli addetti alle ammissioni mirano a reclutare un corpo studentesco con interessi diversi e con la capacità di contribuire alla comunità più ampia.

Inoltre, potrebbero esserci serie sfide pratiche nel misurare il reale risultato degli interessi, lasciando ai decisori il compito di trovare alternative che potrebbero fungere da proxy ragionevole per esso. Nella maggior parte dei casi, i decisori si stabiliscono su un obiettivo di convenienza, ovvero su un obiettivo per il quale esistono dati facilmente accessibili.^{56, 77} Ad esempio, i dati sugli arresti (vale a dire, se qualcuno è stato arrestato) sono spesso adottati come proxy per i dati sui crimini (vale a dire, se qualcuno ha commesso un crimine), anche se molti crimini non vengono mai osservati e quindi non portano mai all'arresto e anche se la polizia potrebbe essere piuttosto selettiva nella scelta di arrestare qualcuno per un crimine osservato.⁷⁸ Senza condonare la decisione di adottare questo obiettivo, potremmo ancora riconoscere le sfide pratiche che incoraggerebbero la polizia a fare affidamento sugli arresti. È semplicemente impossibile osservare tutti i crimini e quindi i decisori potrebbero sentirsi giustificati nel optare per gli arresti come sostituto.

Anche se i decisori avessero qualche modo per ottenere informazioni sulla criminalità, non è ancora ovvio quanto l'obiettivo scelto possa corrispondere agli obiettivi sottostanti della polizia. Prevedere con precisione il verificarsi di crimini futuri non è la stessa cosa che contribuire a ridurre la criminalità; infatti, previsioni accurate sulla criminalità potrebbero semplicemente indurre la polizia a osservare più crimini e a generare più arresti invece di impedire che tali crimini accadano.⁷⁹ Se il vero obiettivo della polizia è ridurre la criminalità e non semplicemente garantire che tutti i crimini portare ad arresti, anche usare la criminalità come bersaglio della previsione potrebbe non aiutare la polizia a realizzare questi obiettivi. Per la polizia potrebbe essere meglio stimare l'effetto deterrente dell'intervento della polizia, ma questo è un compito molto più complicato che fare previsioni sulla base di dati osservativi; rispondere a queste domande richiede sperimentazione. (Naturalmente, anche questa formulazione del problema dovrebbe essere soggetta a ulteriore analisi critica perché non tiene conto dei molti altri tipi di interventi che potrebbero aiutare a ridurre la criminalità oltre a migliorare l'effetto deterrente della presenza della polizia.) Tuttavia, anche quando ci sono buone ragioni per favorire un approccio più sfumato lungo queste linee, i decisori possono favorire semplificazioni imperfette del problema perché sono meno costose o più trattabili.⁶¹

⁵⁶ Infine, i decisori e i soggetti decisionali potrebbero avere idee molto diverse su ciò che costituirebbe il diritto obiettivo della previsione. Gran parte della discussione

In questo capitolo ci si è finora basati sull'idea che gli obiettivi dei decisori sono ampiamente percepiti innanzitutto come desiderabili, e quindi difendibili. Ma ci sono molte volte in cui la questione normativa non riguarda il modo in cui vengono prese le decisioni, ma l'obiettivo del processo decisionale stesso.⁷⁷ In alcuni casi, potremmo non essere d'accordo con gli obiettivi di un dato decisore perché non siamo d'accordo. Non pensare che siano ciò che è nel migliore interesse degli stessi decisori.

Più spesso, potremmo non essere d'accordo con questi obiettivi perché sono in contrasto con gli interessi di altre persone che saranno influenzate negativamente dal perseguitamento di questi obiettivi da parte dei decisori. Come ha sostenuto Oscar Gandy, "certi tipi di bias sono inerenti alla selezione degli obiettivi o delle funzioni obiettivo che i sistemi automatizzati [saranno] progettati per supportare".⁸⁰

Apprezzare come questo sia diverso da un disallineamento target-obiettivo , si consideri un noto studio di Ziad Obermeyer, Brian Powers, Christine Vogeli, et al. sulla parzialità di un algoritmo utilizzato da un sistema sanitario per prevedere quali pazienti trarrebbero maggiori benefici da un programma di "gestione dell'assistenza ad alto rischio".⁸¹ Hanno scoperto che l' algoritmo mostrava pregiudizi razziali – in particolare, che sottostimava il grado in cui la salute dei pazienti neri trarrebbero beneficio dall'iscrizione al programma. Questo perché gli sviluppatori hanno adottato i costi sanitari come obiettivo della previsione, nella convinzione apparente che servisse da indicatore ragionevole dei bisogni sanitari.

Il racconto comune di questa storia suggerisce che i decisori semplicemente non sono riusciti a riconoscere il fatto che ci sono disparità razziali sia nella ricerca di cure che nella fornitura di assistenza sanitaria che fanno sì che i pazienti neri con condizioni di salute altrettanto precarie siano meno costosi rispetto ai pazienti non neri. Pertanto, risolvere il problema richiederebbe solo l'adozione di un obiettivo che rifletta meglio gli obiettivi del sistema sanitario: massimizzare i benefici sanitari complessivi del programma. Eppure è del tutto possibile che l'obiettivo originario della previsione riflettesse i veri obiettivi del sistema sanitario , che avrebbero potuto essere semplicemente quello di ridurre i costi senza alcun riguardo per la salute di chi trarrebbe maggiori benefici da questi interventi. Se così fosse, allora la scelta dell'obiettivo non sarebbe stata semplicemente una scarsa corrispondenza con gli obiettivi dei decisori; gli obiettivi stessi erano problematici. Dobbiamo stare attenti a non confondere i casi in cui ci opponiamo agli obiettivi con i casi in cui ci opponiamo alla particolare scelta dell'obiettivo.

Non prendere in considerazione le informazioni rilevanti

Le burocrazie vengono spesso criticate per non essere sufficientemente individualizzate o particolareggiate nelle loro valutazioni, raggruppando le persone in gruppi inutilmente grossolani. Se i decisori avessero considerato solo qualche dettaglio aggiuntivo, si sarebbero resi conto che la persona in questione è in realtà diversa dal resto delle persone con cui sono stati raggruppati.

L'apprendimento automatico supervisionato è una forma di ragionamento induttivo. Ha lo scopo di trarre regole generali da una serie di esempi specifici, identificando le caratteristiche e i valori delle caratteristiche che coesistono in modo affidabile con un risultato di interesse. A quanto pare, il limite di non essere sufficientemente individualizzati è una parte inevitabile del ragionamento induttivo.

Immaginate una compagnia di assicurazioni automobilistiche che sta cercando di prevedere la probabilità che una persona che richiede una polizza assicurativa subisca un incidente costoso. L'assicuratore cercherà di rispondere a questa domanda guardando la frequenza degli incidenti passati che hanno coinvolto altre persone simili al richiedente. Si tratta di un ragionamento induttivo: è probabile che il richiedente mostri un comportamento simile o sperimenti risultati simili a quelli dei precedenti assicurati perché possiede molte altre qualità in comune con questi assicurati. Forse la persona sta richiedendo un'assicurazione per coprire la propria auto sportiva rosso brillante, un tipo di auto che è coinvolta in incidenti molto più frequentemente rispetto ad altri tipi di auto. Tenendo presente questo andamento storico, l'assicuratore potrebbe quindi concludere che esiste una maggiore possibilità che il richiedente debba presentare un reclamo contro la sua polizza – e offrirsi di assicurare il richiedente solo a un prezzo elevato. Ricevuta l'offerta, il richiedente, che in realtà è un guidatore molto esperto e con un'ottima padronanza dell'auto, potrebbe esitare sul prezzo, obiettando all'idea che presenti un rischio simile a quello degli altri assicurati con la stessa auto .

Cosa è successo qui? L'assicuratore ha fatto la sua previsione sulla base di dettagli piuttosto grossolani (in questo caso solo sul modello e sul colore dell'auto), considerando come indicatore affidabile la velocità con cui si verificano gli incidenti tra i precedenti assicurati con tale auto. della probabilità che il richiedente abbia un incidente. Frederick Schauer si riferisce a questo come al problema delle "generalizzazioni statisticamente valide ma non universali": quando un individuo soddisfa tutti i criteri per l'inclusione in un particolare gruppo, ma non riesce a possedere la qualità che questi criteri dovrebbero prevedere.⁸² Situazioni di questo tipo può dar luogo a denunce di stereotipizzazione o profilazione e alla richiesta che i decisori valutino le persone come individui, non semplicemente come membri di un gruppo. Tuttavia, come ha spiegato Schauer, può essere difficile specificare cosa significhi trattare qualcuno come un

individuo o prendere decisioni individualizzate. Non è chiaro, ad esempio, come un assicuratore possa fare previsioni sulla probabilità di un individuo di rimanere coinvolto in un incidente stradale senza confrontare il richiedente con altre persone che gli somigliano. La questione in questi casi non è il mancato trattamento di qualcuno come individuo, ma la mancata considerazione di ulteriori criteri rilevanti che distinguerebbero una persona dalle altre persone con le quali altrimenti verrebbe confusa.⁸² Se l'assicuratore avesse accesso a ulteriori dettagli (in particolare, dettagli sulle capacità di guida del richiedente e dei precedenti assicurati), l' assicuratore avrebbe potuto esprimere un giudizio più perspicace sul richiedente. Questo è esattamente ciò che accade quando gli assicuratori accettano di offrire prezzi più bassi ai richiedenti che installano volontariamente scatole nere sulle loro auto e che si dimostrano guidatori attenti. È facile interpretare erroneamente questa tendenza come un passo verso una valutazione individualizzata, come se gli assicuratori giudicassero ogni singola persona in base ai suoi meriti unici come conducente. L'interpretazione corretta richiede che si riconosca che gli assicuratori sono in grado di utilizzare i dati della scatola nera di un conducente specifico solo confrontandoli con i dati delle scatole nere di altri conducenti i cui documenti di guida vengono utilizzati per fare una previsione sul conducente. in questione. Anche se accettiamo che le decisioni non possano mai essere completamente individualizzate, potremmo comunque aspettarci che i decisori tengano conto dell'intera gamma di informazioni rilevanti al loro potenziale.

disposizione. Per riprendere l'esempio di cui sopra, potremmo dire che la compagnia di assicurazione auto aveva l'obbligo di considerare le capacità di guida dei richiedenti, e non solo il modello e il colore della loro auto, anche se ciò significava comunque che essi venivano valutati in base a come spesso altre persone con capacità di guida simili e automobili simili hanno avuto incidenti in passato.

Ma fino a che punto dovrebbe estendersi questa aspettativa? Quali obblighi hanno i decisori nel cercare ogni minima informazione immaginabile che potrebbe consentire previsioni più accurate? Bene, a un certo punto le informazioni aggiuntive cessano di essere utili perché non ci sono abbastanza dati di addestramento. Ad esempio, le persone che vivono vicino a un incrocio specifico potrebbero avere maggiori probabilità di avere incidenti perché l'incrocio è mal progettato e quindi pericoloso. Ma l'assicuratore può apprenderlo solo se dispone di dati sufficienti provenienti da un numero sufficiente di persone che vivono vicino a questo incrocio.

C'è anche una ragione molto pratica per cui potremmo non imporre ai decisori uno standard in base al quale sono tenuti a considerare tutte le informazioni che potrebbero essere plausibilmente rilevanti. Raccogliere e considerare tutte queste informazioni può essere costoso, invadente e poco pratico. In effetti, il costo di ciò potrebbe facilmente superare i benefici percepiti che derivano da un processo decisionale più granulare, non solo per il decisore ma anche per i soggetti delle decisioni. Sebbene le scatole nere possano aiutare a raggiungere una maggiore granularità nei prezzi assicurativi, sono anche controverse perché sono piuttosto invadenti e rappresentano una minaccia per la privacy dei conducenti. Per ragioni in questo senso, Schauer e altri hanno suggerito che i decisori sono giustificati nel prendere decisioni sulla base di un insieme limitato di informazioni, anche quando potrebbero esistere ulteriori informazioni rilevanti, se il costo per ottenere tali informazioni ne vanifica i benefici.^{82 , 83} Ci sono tre cose da notare su questi argomenti.

In primo luogo, questi non sono argomenti specifici relativi al processo decisionale automatizzato; sono dichiarazioni generali su qualsiasi forma di processo decisionale, automatizzato o meno. Tuttavia, come abbiamo discusso in precedenza nel capitolo, il processo decisionale automatizzato spesso limita la possibilità di introdurre ulteriori informazioni rilevanti nel processo decisionale. Il risparmio sui costi che potrebbe essere ottenuto automatizzando determinate decisioni (spesso sostituendo i lavoratori umani con software) avviene al costo di privare le persone della possibilità di evidenziare informazioni rilevanti che non trovano posto nel processo automatizzato. Dato che le persone potrebbero essere molto disposte e perfettamente in grado di fornire volontariamente queste informazioni (cioè in grado di farlo a un costo contenuto), il processo decisionale automatizzato che semplicemente nega alle persone l'opportunità di farlo potrebbe fallire l'analisi costi-benefici. In secondo luogo, l'analisi costi-benefici che sostiene le argomentazioni di Schauer e altri non tiene conto di alcuna considerazione distributiva, come quali gruppi potrebbero godere di maggiori benefici o subire maggiori costi. Nel capitolo 4 ritorneremo su questa domanda, chiedendoci se i decisori siano giustificati nel sottoporre determinati gruppi a decisioni meno granulari e quindi meno accurate semplicemente perché ci sono meno informazioni su di loro. Infine, queste argomentazioni non si confrontano con il fatto che i decisori e i soggetti decisionali potrebbero arrivare a conclusioni piuttosto diverse quando eseguono un'analisi costi-benefici se eseguono questa analisi dalla propria prospettiva. Un decisore potrebbe scoprire che i costi per la raccolta di maggiori informazioni non generano un reddito sufficiente

grande beneficio corrispondente per loro come decisori, nonostante il fatto che alcuni soggetti decisionali trarrebbero sicuramente beneficio da un tale investimento. Non è ovvio il motivo per cui si dovrebbe consentire all'analisi costi-benefici dei soli decisori di determinare il livello di granularità accettabile. Una possibile spiegazione potrebbe essere che l'aumento dei costi del processo decisionale (ad esempio, ricercando e prendendo in considerazione maggiori informazioni) incoraggerà i decisori a trasferire semplicemente questi costi sui soggetti decisionali. Ad esempio, se lo sviluppo di una valutazione molto più dettagliata dei richiedenti un'assicurazione auto aumenta i costi operativi dell'assicuratore, è probabile che l'assicuratore addebiti ai richiedenti un prezzo più elevato per compensare questi costi aggiuntivi. Da questa prospettiva, i costi per il decisore sono in realtà solo costi per i soggetti decisionali. Naturalmente, questa prospettiva non contempla la possibilità che l'assicuratore si assuma semplicemente questi costi e accetti meno profitti.

I limiti dell'induzione

Oltre alle considerazioni sui costi, ci sono altri limiti al ragionamento induttivo. Supponiamo che l'allenatore di una squadra di atletica valuti i potenziali membri della squadra in base al colore delle loro scarpe da ginnastica piuttosto che alla velocità con cui possono correre. Immagina che, solo per coincidenza, i corridori più lenti nel pool preferiscano scarpe da ginnastica rosse e i corridori più veloci preferiscano scarpe da ginnastica blu, ma che nessuna relazione simile si ottenga in altri gruppi di corridori. Pertanto, qualsiasi lezione che l'allenatore potrebbe trarre da questi particolari corridori sulla relazione tra i colori delle scarpe da ginnastica e la velocità sarebbe inaffidabile se applicata ad altri corridori. Questo è il problema del sovraccarico.84 Si tratta di una forma di processo decisionale arbitrario perché la validità predittiva che funge da giustificazione è un'illusione.

L'overfitting è un problema ben noto nell'apprendimento automatico ed esistono molti modi per contrastarlo. Poiché la relazione spuria si verifica a causa di una coincidenza, quanto più grande è il campione, tanto meno probabile è che si verifichi. Inoltre, è possibile penalizzare i modelli eccessivamente complicati per rendere meno probabile che riconoscano modelli casuali nei dati. E, cosa più importante, è pratica standard separare gli esempi utilizzati per addestrare e testare i modelli di machine learning. Ciò consente una valutazione realistica della capacità delle relazioni osservate nei dati di addestramento di essere trasferite ad esempi mai visti. Per questi motivi, a meno che non si tratti di campioni di piccole dimensioni, il sovraccarico generalmente non è un problema serio nella pratica.

Ma le varianti del problema del sovraccarico possono essere molto più gravi e spinose. È pratica comune nell'apprendimento automatico prendere un set di dati esistente, in cui tutti i dati sono stati raccolti in modo simile, e dividere semplicemente questo set di dati in set di training e test. Le piccole differenze tra questi set aiuteranno a evitare l'adattamento eccessivo e potrebbero dare un senso alle prestazioni su dati invisibili. Ma queste divisioni sono ancora molto più simili tra loro rispetto alla futura popolazione a cui il modello potrebbe essere applicato.85,86 Questo è il problema dei "cambiamenti distributivi", di cui esistono molti tipi diversi. Sono comuni nella pratica e rappresentano un problema fondamentale per il paradigma dell'apprendimento automatico.

Tornando al nostro esempio precedente, immagina che i corridori possano solo acquistare

scarpe da ginnastica di un fornitore e che il fornitore vende solo un tipo di scarpe da ginnastica, ma varia il colore della scarpa in base alla taglia (tutte le taglie inferiori alla 8 sono rosse, mentre tutte le taglie dalla 8 in su sono blu). Inoltre, supponiamo che i corridori con i piedi più grandi siano più veloci di quelli con i piedi più piccoli e che vi sia un grande cambiamento di passo nella velocità dei corridori una volta che la misura del loro piede supera 7. In queste circostanze, la selezione dei corridori in base al colore delle loro scarpe da ginnastica risulterà attendibilmente in una squadra composta da corridori più veloci, ma lo farà per ragioni che potremmo comunque trovare insensate o addirittura discutibili. Perché? La relazione tra il colore delle scarpe da ginnastica di un corridore e la velocità di corsa è ovviamente spuria, nel senso che sappiamo che il colore delle scarpe da ginnastica di un corridore non ha alcun effetto causale sulla velocità. Ma questa relazione è davvero spuria? Non è semplicemente un artefatto del particolare insieme di esempi da cui è stata indotta una regola generale; è una relazione stabile nel mondo reale. Finché rimane un solo fornitore e il fornitore offre solo colori diversi in queste taglie specifiche, il colore delle scarpe distinguerà in modo affidabile i corridori più veloci da quelli più lenti. Allora qual è il problema nel prendere decisioni su questa base? Ebbene, potremmo non avere sempre un modo per determinare se stiamo operando nelle condizioni descritte. Generalizzare partendo da esempi specifici ammette sempre la possibilità di trarre lezioni che non si applicano alla situazione che i decisori dovranno affrontare in futuro.

Una risposta a queste preoccupazioni è affermare che esiste un obbligo normativo secondo cui i criteri decisionali hanno una relazione causale con il risultato che vengono utilizzati per prevedere. Il problema con l'utilizzo del colore delle scarpe da ginnastica come criterio è per noi evidente perché riconosciamo la completa assenza di qualsiasi plausibile influenza causale sulla velocità di corsa. Quando si utilizza l'apprendimento automatico, i modelli risultanti, indifferenti alla causalità, possono sfruttare correlazioni instabili.⁸⁷ Ciò fa sorgere la richiesta che nessuno sia soggetto a schemi decisionali basati su risultati privi di valore scientifico – vale a dire, su risultati spuri e quindi non validi. Probabilmente spiegano le preoccupazioni di studiosi come Frank Pasquale, che parla di casi in cui l'apprendimento automatico è “faccialmente non valido”,⁸⁸ e Pauline Kim ed Erika Hanson, che hanno sostenuto che “poiché il data mining scopre relazioni statistiche che potrebbero non essere causali, basandosi su tali correlazioni per fare previsioni su casi futuri può portare a un trattamento arbitrario degli individui”.⁸⁹ Affermare che gli schemi decisionali dovrebbero essere basati solo su criteri che hanno una relazione causale con l'esito di interesse è probabilmente percepito come un modo per evitare questi situazioni – cioè come un modo per garantire che la base per il processo decisionale sia ben ragionata e non arbitraria.

Il diritto a previsioni accurate?

Nelle due sezioni precedenti, abbiamo discusso diversi motivi per cui le previsioni che utilizzano il ragionamento induttivo potrebbero essere imprecise, inclusa la mancata considerazione delle informazioni rilevanti e lo spostamento della distribuzione. Ma anche se mettiamo da parte queste ragioni – supponiamo che il decisore consideri tutte le informazioni disponibili, non vi sia alcun cambiamento nella distribuzione, ecc. – potrebbero esserci limiti insormontabili all'accuratezza della previsione.

risultati futuri. Questi limiti potrebbero persistere indipendentemente dal fatto che venga utilizzato o meno il ragionamento induttivo.⁹⁰ Ad esempio, almeno alcuni casi di recidiva sono dovuti a crimini improvvisati commessi quando si sono presentate fortuitamente opportunità e questi potrebbero non essere prevedibili in anticipo. (Nei capitoli successivi esamineremo alcune delle prove empiriche dei limiti alla previsione.)

Quali sono le implicazioni di questi limiti alla previsione? Dal punto di vista del decisore, anche un piccolo aumento dell'accuratezza predittiva rispetto a uno scenario di base (giudizio umano o politica basata su regole) può essere prezioso. Consideriamo un'agenzia per la protezione dell'infanzia che utilizzi uno strumento di screening predittivo per determinare quali bambini sono a rischio di abusi sui minori. Una maggiore precisione può significare un minor numero di bambini affidati in affidamento. Potrebbe anche comportare un notevole risparmio sui costi, con un minor numero di operatori sociali necessari per effettuare visite a domicilio.

Un modello tipico utilizzato nella pratica può avere una precisione (più precisamente, AUC) compresa tra 0,7 e 0,8.⁹¹ È meglio del lancio di una moneta, ma comporta comunque un numero considerevole di falsi positivi e falsi negativi. L'affermazione che il sistema rende possibile la decisione più accurata al momento dello screening è un magro conforto per le famiglie in cui i bambini sono separati dai genitori a causa della previsione del modello di futuri abusi, o per i casi di abuso che il modello prevedeva essere a basso rischio.

Se i risultati del modello fossero casuali, lo considereremmo chiaramente arbitrario e illegitimo (e persino crudele). Ma qual è la soglia di precisione per la legittimità? In altre parole, quanto deve essere elevata la precisione per poter giustificare l'uso di un sistema predittivo?⁹²

La bassa accuratezza diventa ancora più problematica se consideriamo che viene misurata rispetto a un obiettivo di previsione che in genere richiede il sacrificio di alcuni degli obiettivi sfaccettati che i decisori potrebbero avere. Ad esempio, un modello di previsione del rischio per il benessere dei bambini potrebbe non essere in grado di ragionare sugli effetti differenziali che un intervento come l'affidamento potrebbe avere su bambini e famiglie diverse. Quanto aumento dell'accuratezza predittiva è necessario per giustificare la discrepanza tra gli obiettivi effettivi del sistema e quelli realizzati dal modello?

Ovviamente, queste domande non hanno risposte facili, ma rappresentano minacce importanti e sottovalutate alla legittimità del processo decisionale predittivo.

Agenzia, ricorso e colpevolezza

Consideriamo ora una questione molto diversa: i criteri che mostrano rilevanza statistica e consentono previsioni accurate potrebbero ancora essere normativamente inappropriati come base per il processo decisionale?

Forse il criterio in questione è una caratteristica immutabile. Forse è una caratteristica mutevole, ma non una caratteristica che la persona specifica in questione abbia la capacità di cambiare. O forse la caratteristica è stata influenzata dalle azioni di altri e non è il risultato delle azioni della persona. Ognuna di queste ragioni, in modi leggermente diversi, riguarda tutte il grado di controllo che si ritiene che una persona abbia sulla caratteristica in questione - e ciascuna fornisce una qualche giustificazione normativa per ignorare o sminuire la caratteristica anche se

quando potrebbe essere dimostrabilmente predittivo del risultato di interesse. Esaminiamo ulteriormente ciascuna di queste preoccupazioni.

Le decisioni basate su caratteristiche immutabili possono essere motivo di preoccupazione perché minacciano l'azione delle persone. Per definizione, non c'è niente che qualcuno possa fare per cambiare caratteristiche immutabili (ad esempio, il proprio paese di nascita). Per estensione, non c'è niente che nessuno possa fare per cambiare le decisioni prese sulla base di caratteristiche immutabili. In queste circostanze, le persone sono condannate al proprio destino e non sono più protagoniste della propria vita. C'è qualcosa di inquietante nell'idea di privare le persone della capacità di apportare cambiamenti che porterebbero a un risultato diverso dal processo decisionale, soprattutto quando queste decisioni potrebbero influenzare in modo significativo le possibilità di vita e il corso della vita di una persona. Ciò potrebbe essere considerato particolarmente problematico quando sembrano esserci modi alternativi in cui un decisore può esprimere un giudizio efficace su una persona senza fare affidamento su caratteristiche immutabili. In questa prospettiva, se è possibile sviluppare schemi decisionali che siano ugualmente accurati, ma che lascino comunque spazio ai soggetti decisionali per adattare il loro comportamento in modo da aumentare le loro possibilità di una decisione favorevole, allora i decisori hanno l'obbligo morale di adottare un tale schema per rispetto dell'azione delle persone.

Il ricorso è un'idea correlata ma più generale relativa al grado in cui le persone hanno la capacità di apportare cambiamenti che si traducono in decisioni diverse.⁹³ Sebbene non vi sia nulla che qualcuno possa fare per cambiare una caratteristica immutabile, le persone potrebbero essere più o meno capaci di cambiarle. caratteristiche che sono, in linea di principio, mutevoli.^{94, 95} Alcune persone potrebbero aver bisogno di spendere molte più risorse di altre per ottenere il risultato che desiderano dal processo decisionale. La scelta di determinati criteri da utilizzare come base per il processo decisionale è anche una scelta sul tipo di azioni che le persone potranno intraprendere nel cercare una decisione diversa. E le persone in circostanze diverse avranno capacità diverse per farlo con successo. In alcuni casi, le persone potrebbero non avere mai risorse sufficienti per raggiungere questo obiettivo, riportandoci alla stessa situazione discussa nel paragrafo precedente. Ad esempio, un richiedente di credito potrebbe essere ben posizionato per trasferirsi in un nuovo quartiere in modo da diventare un candidato più attraente per un nuovo prestito, presupponendo che lo schema decisionale utilizzi l'ubicazione come criterio importante. Ma un altro candidato potrebbe non essere in grado di farlo, per ragioni finanziarie, culturali o per molti altri motivi.

La ricerca sul ricorso all'apprendimento automatico si è concentrata in gran parte sull'assicurare che le persone ricevano spiegazioni sui modi per raggiungere una decisione diversa da un modello che le persone possono effettivamente eseguire nella realtà.⁹⁶ Dato che ci sono molti modi possibili per spiegare le decisioni di un modello di apprendimento automatico , l'obiettivo di questo lavoro è quello di garantire che le spiegazioni offerte indirizzino le persone a intraprendere azioni fattibili piuttosto che suggerire che l'unico modo per ottenere il risultato desiderato sia fare qualcosa che va oltre le loro capacità. Anche quando si sviluppa uno schema decisionale che si basa solo su caratteristiche mutevoli, i decisori possono fare di più per preservare il ricorso, adattando la loro spiegazione delle decisioni di un modello per concentrarsi su quelle azioni che sono più facili da cambiare per le persone. Per questo motivo, quanto più le persone sono in grado di apportare modifiche che diano loro il risultato desiderato, migliore sarà il processo decisionale

schema e migliore sarà la spiegazione.

Infine, come accennato in precedenza in questa sezione, potremmo considerare ingiusti alcuni schemi decisionali se ritengono le persone responsabili di caratteristiche al di fuori del loro controllo. Le idee di base sulla colpevolezza morale si basano quasi sempre su una certa comprensione delle azioni che hanno prodotto gli esiti preoccupanti.

Ad esempio, potremmo essere arrabbiati con una persona che ci ha urtato facendoci cadere e rompendo un oggetto prezioso. Quando scopriamo che sono stati spinti da qualcun altro, è probabile che li riterremo incolpevoli e reindirizzeremo la nostra disapprovazione alla persona che li ha spinti. Questo stesso ragionamento spesso si ripercuote sul modo in cui pensiamo all'equità di fare affidamento su determinati criteri quando prendiamo decisioni che allocano risorse e opportunità desiderabili. A meno che non sappiamo perché certi risultati si avverano, non possiamo giudicare se i decisori siano giuridicamente giustificati nel fare affidamento su criteri che predicono accuratamente se quel risultato si realizzerà. Dobbiamo comprendere la causa dell'esito di interesse in modo da poter riflettere se l'oggetto della decisione abbia la responsabilità morale per l'esito, data la sua causa.

Ad esempio, come ha esplorato Barbara Kiviat, le leggi in molti stati degli Stati Uniti limitano il grado in cui i fornitori di assicurazioni auto possono prendere in considerazione "circostanze di vita straordinarie" quando prendono decisioni di sottoscrizione o di prezzo, inclusi eventi come la morte del coniuge, di un figlio, di un o un genitore.⁹⁷ Queste leggi vietano agli assicuratori di prendere in considerazione una serie di fattori sui quali le persone non possono esercitare alcun controllo – come un decesso in famiglia – ma che potrebbero comunque contribuire a far sì che qualcuno si trovi in difficoltà finanziarie e quindi ad aumentare la probabilità di presentare una richiesta di risarcimento nei confronti di la loro polizza di assicurazione auto anche in caso di incidente lieve. Questi divieti riflettono la convinzione di fondo secondo cui le persone non dovrebbero essere soggette a decisioni avverse se non fossero responsabili di qualunque cosa le faccia apparire meno meritevoli di un trattamento più favorevole. O per dirla in un altro modo: le persone dovrebbero essere giudicate solo sulla base di fattori per i quali sono moralmente colpevoli. La piena attuazione di questo principio non è pratica, poiché la maggior parte degli attributi che il decisore potrebbe utilizzare, ad esempio il reddito, sono in parte ma non completamente il risultato delle scelte dell'individuo. Tuttavia, attributi come la morte in famiglia sembrano ricadere abbastanza chiaramente su un lato della linea.

Naturalmente, c'è un rovescio della medaglia in tutto questo. Se le persone potessero modificare facilmente le funzionalità utilizzate per prendere decisioni su di loro, potrebbero "ingannare" il processo决策的. Per gioco intendiamo cambiare il valore delle caratteristiche al fine di cambiare la decisione senza cambiare il risultato atteso che le caratteristiche dovrebbero prevedere.⁹⁸ "Insegnare per mettere alla prova" è uno scenario familiare che è un esempio di gioco. In questo caso, il punteggio del test è una caratteristica che prevede le prestazioni future (ad esempio, in un lavoro). Supponiamo che il test, in effetti, abbia un valore predittivo, perché le persone che ottengono buoni risultati nel test tendono ad aver padroneggiato alcune conoscenze sottostanti, e tale padronanza migliora le loro prestazioni lavorative. L'insegnamento del test si riferisce a metodi di preparazione che aumentano il punteggio del test senza aumentare corrispondentemente l'abilità di base che il punteggio dovrebbe riflettere.

Ad esempio, gli insegnanti potrebbero aiutare gli studenti a prepararsi per il test sfruttando il fatto che il test valuta conoscenze o competenze molto specifiche, non l'intero

gamma di conoscenze o competenze che il test intende misurare e concentrare la preparazione solo su quelle parti che verranno valutate.⁹⁹ Jane Bambauer e Tal Zarsky forniscono molti esempi di sistemi decisionali nel gioco.¹⁰⁰

Il gioco è un problema comune perché la maggior parte dei modelli non scopre il meccanismo causale che spiega il risultato. Pertanto, prevenire il gioco richiede un modello causale.⁹⁸ Inoltre, uno schema giocabile diventa meno efficace nel tempo e può compromettere gli obiettivi del decisore e la corretta allocazione delle risorse. In effetti, il gioco può rappresentare un problema anche quando i soggetti decisionali non agiscono in modo avversario. Le persone in cerca di lavoro possono spendere considerevoli sforzi e denaro per ottenere credenziali senza senso che gli viene detto siano importanti nel loro settore, solo per scoprire che, sebbene questo li aiuti a trovare un lavoro, non li rende più preparati a svolgerlo effettivamente.¹⁰¹ In tali circostanze, il comportamento strategico può rappresentare uno spreco di sforzi da parte di attori ben intenzionati.

Considerazioni conclusive

In questo capitolo abbiamo analizzato tre forme di automazione. Abbiamo discusso di come ciascuno di questi risponda in qualche modo alle preoccupazioni relative al processo decisionale arbitrario, apprendo allo stesso tempo nuove preoccupazioni sulla legittimità. Abbiamo poi approfondito il terzo tipo di automazione, l'ottimizzazione predittiva, che è ciò di cui ci occuperemo nella maggior parte di questo libro.

Per essere chiari, non facciamo affermazioni generali sulla legittimità del processo decisionale automatizzato o dell'ottimizzazione predittiva. Nelle applicazioni che non sono consequenziali alle possibilità di vita delle persone, le questioni di legittimità sono meno importanti. Ad esempio, nel rilevamento delle frodi con carte di credito, vengono utilizzati modelli statistici per individuare modelli nei dati delle transazioni, come un improvviso cambiamento di posizione, che potrebbero indicare una frode derivante dal furto dei dati della carta di credito. La posta in gioco per i singoli individui tende ad essere piuttosto bassa. Negli Stati Uniti, ad esempio, la responsabilità individuale è limitata a 50 dollari a condizione che vengano soddisfatte determinate condizioni. Pertanto, sebbene gli errori siano costosi, il costo è sostenuto principalmente da chi prende le decisioni (in questo esempio, la banca). Pertanto le banche tendono a implementare tali modelli sulla base di considerazioni sui costi senza preoccuparsi (ad esempio) di fornire ai clienti un modo per contestare il modello.

Nelle applicazioni consequenziali, tuttavia, per stabilire la legittimità, i decisori devono essere in grado di giustificare affermativamente il loro schema lungo le dimensioni che abbiamo delineato: spiegare come l'obiettivo si collega agli obiettivi su cui tutte le parti interessate possono concordare; convalidare l'accuratezza del sistema distribuito; consentire metodi di ricorso e così via. In molti casi, è possibile mettere protezioni procedurali attorno ai sistemi automatizzati per ottenere questa giustificazione, ma i decisori sono restii a farlo perché ciò compromette i risparmi sui costi che l'automazione dovrebbe ottenere.

3

Classificazione

L'obiettivo della classificazione è sfruttare i modelli nei processi naturali e sociali per fare congetture su risultati incerti. Un risultato può essere incerto perché si trova nel futuro. Questo è il caso quando proviamo a prevedere se un richiedente rimborsa un prestito esaminando varie caratteristiche come la storia creditizia e il reddito. La classificazione si applica anche a situazioni in cui l'esito si è già verificato, ma non ne siamo sicuri. Ad esempio, potremmo provare a classificare se si è verificata una frode finanziaria esaminando le transazioni finanziarie.

Ciò che rende possibile la classificazione è l'esistenza di modelli che collegano l'esito di interesse in una popolazione a informazioni che possiamo osservare. La classificazione è specifica per una popolazione e per i modelli prevalenti nella popolazione. I richiedenti di prestiti rischiosi potrebbero avere un track record di elevato utilizzo del credito. La frode finanziaria spesso coincide con irregolarità nella distribuzione delle cifre nei rendiconti finanziari. Questi modelli potrebbero esistere in alcuni contesti ma non in altri. Di conseguenza, il grado in cui funziona la classificazione varia.

Formalizziamo la classificazione in due passaggi. Il primo è rappresentare una popolazione come distribuzione di probabilità. Anche se oggi viene spesso dato per scontato nel lavoro quantitativo, l'atto di rappresentare una popolazione dinamica di individui come una distribuzione di probabilità rappresenta un cambiamento significativo di prospettiva. Il secondo passo consiste nell'applicare la statistica, in particolare la teoria delle decisioni statistiche, alla distribuzione di probabilità che rappresenta la popolazione. La teoria delle decisioni statistiche formalizza l'obiettivo della classificazione, permettendoci di parlare della qualità dei diversi classificatori.

Il trattamento teorico-decisionale statistico della classificazione costituisce il fondamento dell'apprendimento automatico supervisionato. L'apprendimento supervisionato rende la classificazione algoritmica nel modo in cui fornisce euristiche per trasformare i campioni di una popolazione in buone regole di classificazione.

Modellare le popolazioni come distribuzioni di probabilità

Una delle prime applicazioni della probabilità allo studio delle popolazioni umane è la tavola di vita di Halley del 1693. Halley tabulava le nascite e le morti in una piccola città per stimare l'aspettativa di vita della popolazione. Le stime dell'aspettativa di vita, allora innovative quanto la stessa teoria della probabilità, trovarono impiego nel valutare accuratamente gli investimenti che pagavano una somma di denaro ogni anno per il resto di un periodo.

Age.	Per- sons	Age.	Per- sons										
Curt.		Curt.											
1	1000	8	680	15	628	22	585	29	539	36	481	7	5547
2	855	9	670	16	622	23	579	30	531	37	472	14	4584
3	798	10	661	17	616	24	573	31	523	38	463	21	4270
4	750	11	653	18	610	25	567	32	515	39	454	28	3564
5	732	12	646	19	604	26	560	33	507	40	445	35	3604
6	710	13	640	20	598	27	553	34	499	41	436	42	3178
7	692	14	634	21	592	28	546	35	490	42	427	49	2709
												56	2194
Age.	Per- sons	Age.	Per- sons										
Curt.		Curt.		Curt.		Curt.		Curt.		Curt.		Curt.	
43	417	50	346	57	272	64	202	71	131	78	58	77	692
44	407	51	335	58	262	65	192	72	120	79	49	84	253
45	397	52	324	59	252	65	182	73	109	80	41	100	107
46	387	53	313	60	242	67	172	74	98	81	34		
47	377	54	302	61	232	68	162	75	88	82	28		34000
48	367	55	292	62	222	69	152	76	78	83	23		
49	357	56	282	63	212	70	142	77	68	84	20		
												Sum Total.	

Figura 3.1: Tavola della vita di Halley (1693)

la vita della persona.

Nei secoli che seguirono, l'uso della probabilità per modellare le popolazioni umane, tuttavia, rimase controverso sia dal punto di vista scientifico che politico.^{102, 103, 104} Tra i primi ad applicare la statistica alle scienze sociali fu il diciannovesimo astronomo e sociologo Adolphe Quetelet. In un programma scientifico che chiamò "fisica sociale", Quetelet cercò di dimostrare l'esistenza di leggi statistiche nelle popolazioni umane. Ha introdotto il concetto di "uomo medio" caratterizzato dai valori medi delle variabili misurate, come l'altezza, che seguivano una distribuzione normale. Proposta tanto descrittiva quanto normativa, Quetelet considerava le medie come un ideale da perseguire. Tra gli altri, il suo lavoro influenzò Francis Galton nello sviluppo dell'eugenetica.

Il successo della statistica nel corso del XX secolo ha consolidato l'uso della probabilità per modellare le popolazioni umane. Pochi oggi alzano il sopracciglio se parliamo di un sondaggio come di un campionamento delle risposte di una distribuzione. Sembra ovvio ora che vorremmo stimare parametri come la media e la deviazione standard dalle distribuzioni dei redditi, dalle dimensioni delle famiglie o altri attributi simili. La statistica è così profondamente radicata nelle scienze sociali che raramente rivisitiamo la premessa secondo cui possiamo rappresentare una popolazione umana come una distribuzione di probabilità.

Le differenze tra una popolazione umana e una distribuzione sono evidenti. Le popolazioni umane cambiano nel tempo, a volte rapidamente, a causa di azioni, meccanismi e interazioni diverse tra gli individui. Una distribuzione, al contrario, può essere pensata come un array statico in cui le righe corrispondono agli individui e le colonne corrispondono alle covariate misurate di un individuo. L'astrazione matematica per un tale array è un insieme di numeri non negativi, chiamati probabilità, che si sommano a 1 e ci danno per ogni riga il peso relativo di questo insieme di covariate nella popolazione. Campionare da una tale distribuzione equivale a scegliere a caso una delle righe della tabella in proporzione al suo peso. Possiamo ripetere questo processo senza cambiamenti o deterioramenti. In questa visione la distribuzione è immutabile.

Niente di ciò che facciamo può cambiare la popolazione.

Gran parte della statistica riguarda i campioni e la questione di come possiamo mettere in relazione le quantità calcolate su un campione, come la media campionaria, con i parametri corrispondenti di una distribuzione, come la media della popolazione. Il focus nel nostro capitolo è diverso. Utilizzeremo la statistica per parlare delle proprietà delle popolazioni come distribuzioni e, per estensione, delle regole di classificazione applicate a una popolazione. Sebbene il campionamento introduca molte questioni aggiuntive, le domande che solleviamo in questo capitolo emergono più chiaramente a livello di popolazione.

Classificazione formalizzante

L'obiettivo della classificazione è determinare un valore plausibile per un obiettivo Y sconosciuto date le covariate X osservate. In genere, le covariate sono rappresentate come una matrice di variabili continue o discrete, mentre l'obiettivo è un valore discreto, spesso binario.

Formalmente, le covariate X e l'obiettivo Y sono variabili casuali distribuite congiuntamente. Ciò significa che esiste una distribuzione di probabilità su coppie di valori (x, y) che le variabili casuali (X, Y) potrebbero assumere. Questa distribuzione di probabilità modella una popolazione di istanze del problema di classificazione. Nella maggior parte dei nostri esempi, pensiamo a ciascuna istanza come alle covariate e all'obiettivo di un individuo.

Al momento della classificazione, il valore della variabile target non ci è noto, ma osserviamo le covariate X e facciamo un'ipotesi $Y = f(X)$ in base a ciò che abbiamo osservato. La funzione f che mappa le nostre covariate nella nostra ipotesi Y è chiamata classificatore o predittore. L'output del classificatore è chiamato etichetta o previsione.

In questo capitolo siamo interessati principalmente alla variabile casuale Y e alla sua relazione con altre variabili casuali. La funzione che definisce queste variabili casuali è secondaria. Per questo motivo allunghiamo leggermente la terminologia e ci riferiamo a Y stesso come classificatore.

Implicito in questa impostazione formale di classificazione è un presupposto importante. Qualunque cosa facciamo sulla base delle covariate X non può influenzare il risultato Y . Dopotutto, la nostra distribuzione assegna un peso fisso a ciascuna coppia (x, y) . In particolare, la nostra previsione Y non può influenzare il risultato Y . Questo presupposto viene spesso violato quando le previsioni motivano azioni che influenzano il risultato. Ad esempio, la previsione che uno studente sia a rischio di abbandono scolastico potrebbe essere seguita da interventi educativi che rendano meno probabile l'abbandono scolastico.

Per poter scegliere un classificatore tra molte possibilità, dobbiamo formalizzare ciò che rende valido un classificatore. Questa domanda spesso non ha una risposta pienamente soddisfacente, ma la teoria delle decisioni statistiche fornisce criteri che possono aiutare a evidenziare diverse qualità di un classificatore che possono informare la nostra scelta.

Forse la proprietà più nota di un classificatore Y è la sua accuratezza di classificazione, o accuratezza in breve, definita come $P\{Y = Y\}$, la probabilità di prevedere correttamente la variabile target. Definiamo l'errore di classificazione come $P\{Y \neq Y\}$. L'accuratezza è facile da definire, ma trascura alcuni aspetti importanti quando si valuta un classificatore. Un classificatore che preveda sempre l'assenza di incidenti stradali mortali nel prossimo anno potrebbe avere un'elevata precisione su un dato individuo, semplicemente perché gli incidenti mortali sono improbabili.

Tuttavia, è una funzione costante che non ha alcun valore nella valutazione del rischio di un incidente stradale.

Altri criteri teorici della decisione evidenziano diversi aspetti di un classificatore. Possiamo definire quelli più comuni considerando la probabilità condizionata $P\{\text{evento} | \text{condizione}\}$ per varie impostazioni diverse.

Tabella 3.1: Criteri comuni di classificazione

Evento Condizione Nozione risultante ($P\{\text{evento} \text{condizione}\}$)	
$Y = 1$	$Y = 1$
$Y = 0$	$Y = 1$
$Y = 1$	$Y = 0$
$Y = 0$	$Y = 0$
	Tasso di vero positivo, ricorda
	Tasso di falsi negativi
	Tasso di falsi positivi
	Tasso vero negativo

Il vero tasso di positività corrisponde alla frequenza con cui il classificatore assegna correttamente un'etichetta positiva quando il risultato è positivo. Lo chiamiamo un vero positivo. Gli altri termini falso positivo, falso negativo e vero negativo derivano analogamente dalle rispettive definizioni. Non è importante memorizzare tutti questi termini. Tuttavia, compaiono regolarmente nelle impostazioni di classificazione.

Un'altra famiglia di criteri di classificazione nasce dallo scambio di evento e condizione . Evidenzieremo solo due delle quattro possibili nozioni.

Tabella 3.2: Criteri aggiuntivi di classificazione

Evento Condizione Nozione risultante ($P\{\text{evento} \text{condizione}\}$)	
$Y = 1$	$Y = 1$
$Y = 0$	$Y = 0$
	Valore predittivo positivo, precisione
	Valore predittivo negativo

Classificazione ottimale

Supponiamo di assegnare un costo (o ricompensa) quantificato a ciascuno dei quattro possibili risultati della classificazione, vero positivo, falso positivo, vero negativo, falso negativo. Il problema della classificazione ottimale è trovare un classificatore che riduca al minimo i costi attesi su una popolazione. Possiamo scrivere il costo come un numero reale (y, y) , chiamato perdita, che sperimentiamo quando classifichiamo un valore target y con un'etichetta y . Un classificatore ottimale è qualsiasi classificatore che minimizza la perdita attesa:

$$E[(Y, Y)]$$

Questo obiettivo è chiamato rischio di classificazione e la minimizzazione del rischio si riferisce al problema di ottimizzazione di trovare un classificatore che minimizzi il rischio.

Ad esempio, scegli le perdite $(0, 1) = (1, 0) = 1$ e $(1, 1) = (0, 0) = 0$. Per questa scelta della funzione di perdita, il classificatore ottimale è quello che minimizza l'errore di classificazione. Il classificatore ottimale risultante ha una soluzione intuitiva.

Fatto 1. Il predittore ottimale che minimizza l'errore di classificazione è soddisfatto

$$Y = f(X), \text{ dove } f(x) = \begin{cases} 1 & \text{se } P\{Y = 1 | X = x\} > 1/2 \\ 0 & \text{altrimenti.} \end{cases}$$

Il classificatore ottimale controlla se la propensione ai risultati positivi date le covariate X osservate è maggiore di 1/2. In tal caso, suppone che il risultato sia 1. Altrimenti, suppone che il risultato sia 0. Il predittore ottimale sopra è specifico per l'errore di classificazione. Se la nostra funzione di perdita fosse diversa, la soglia 1/2 nella definizione di cui sopra dovrebbe cambiare. Ciò ha un senso intuitivo. Se il nostro costo per i falsi positivi fosse molto più alto del nostro costo per i falsi negativi, faremmo meglio a sbagliare per non dichiarare un positivo.

Il predittore ottimale è una costruzione teorica che potremmo non essere in grado di costruire partendo dai dati. Ad esempio, quando il vettore delle covariate X è ad alta dimensione, un campione finito probabilmente perderà alcune impostazioni $X = x$ che le covariate potrebbero assumere. In questo caso non è chiaro come ottenere la probabilità $P\{Y = 1 | X = x\}$. Esiste un vasto repertorio tecnico in statistica e nell'apprendimento automatico per trovare buoni predittori da campioni finiti. In questo capitolo ci concentreremo sui problemi che persistono anche se avessimo accesso al predittore ottimale per una data popolazione.

Punteggi di rischio

Il classificatore ottimo che abbiamo appena visto possiede una proprietà importante. Siamo riusciti a scriverlo come soglia applicata alla funzione

$$r(x) = P\{Y = 1 | X = x\} = E[Y | X = x].$$

Questa funzione è un esempio di punteggio di rischio. La teoria delle decisioni statistiche ci dice che i classificatori ottimali possono generalmente essere scritti come una soglia applicata a questo punteggio di rischio.

Il punteggio di rischio che vediamo qui è particolarmente importante e naturale. Possiamo pensarlo come se prendessimo le prove disponibili $X = x$ e calcolassimo il risultato atteso date le informazioni osservate. Questa è chiamata probabilità a posteriori del risultato Y dato X . In senso intuitivo, l'aspettativa condizionale è una tabella di ricerca statistica che ci fornisce per ogni impostazione di caratteristiche la frequenza dei risultati positivi date queste caratteristiche. Il punteggio di rischio è talvolta chiamato ottimale di Bayes. Minimizza la perdita al quadrato

$$E(Y - r(X))^2$$

tra tutti i possibili punteggi di rischio a valore reale $r(X)$. I problemi di minimizzazione in cui si tenta di approssimare la variabile target Y con un punteggio di rischio a valore reale sono chiamati problemi di regressione. In questo contesto, i punteggi di rischio sono spesso chiamati regressori. Sebbene la nostra funzione di perdita fosse specifica, esiste una lezione generale. La classificazione viene spesso attaccata risolvendo prima un problema di regressione per riassumere i dati in un unico punteggio di rischio a valore reale. Trasformiamo quindi il punteggio di rischio in un classificatore mediante soglia.

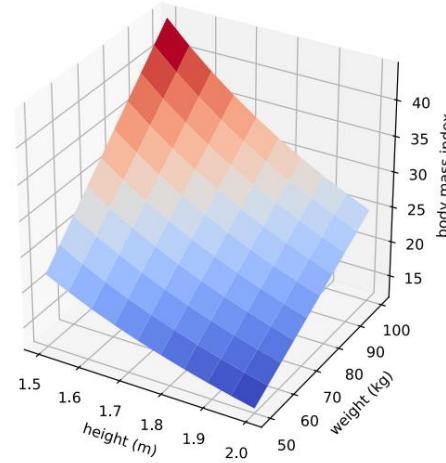


Figura 3.2: Grafico dell'indice di massa corporea.

Non è necessario che i punteggi di rischio siano ottimali o ricavati dai dati. Come esempio illustrativo consideriamo il noto indice di massa corporea, dovuto tra l'altro a Quetelet, che riassume il peso e l'altezza di una persona in un unico numero reale. Nella nostra notazione formale, le caratteristiche sono $X = (H, W)$ dove H indica l'altezza in metri e W indica il peso in chilogrammi. L'indice di massa corporea corrisponde alla funzione di punteggio $R = W/H^2$.

Potremmo interpretare l'indice di massa corporea come una misura del rischio, ad esempio, del diabete. Sogliandolo al valore 30, potremmo decidere che gli individui con un indice di massa corporea superiore a questo valore sono a rischio di sviluppare il diabete mentre altri no. Non è necessaria una laurea in medicina per preoccuparsi che il classificatore risultante possa non essere molto accurato. L'indice di massa corporea presenta una serie di problemi noti che portano a errori quando utilizzato per la classificazione. Non entreremo nei dettagli, ma vale la pena notare che questi errori di classificazione possono allinearsi sistematicamente con determinati gruppi della popolazione. Ad esempio, l'indice di massa corporea tende ad essere gonfiato come misura di rischio per le persone più alte a causa di problemi di ridimensionamento.

Un approccio più raffinato per trovare un punteggio di rischio per il diabete sarebbe quello di risolvere un problema di regressione che coinvolga le covariate disponibili e la variabile di risultato. Risolti in modo ottimale, il punteggio di rischio risultante ci direbbe per ogni impostazione di peso (ad esempio, arrotondato all'unità di kg più vicina) e per ogni altezza fisica (arrotondata all'unità di cm più vicina), il tasso di incidenza del diabete tra gli individui con questi valori di peso e altezza. La variabile target in questo caso è un indicatore binario del diabete. Quindi, $r((176, 68))$ sarebbe il tasso di incidenza del diabete tra gli individui alti 1,76 m pesano 68 kg. L'aspettativa condizionale è probabilmente più utile come misura del rischio di diabete rispetto all'indice di massa corporea che abbiamo visto in precedenza. Dopotutto, l'aspettativa condizionata riflette direttamente il tasso di incidenza del diabete date le caratteristiche osservate, mentre l'indice di massa corporea non ha risolto questo specifico problema di regressione.

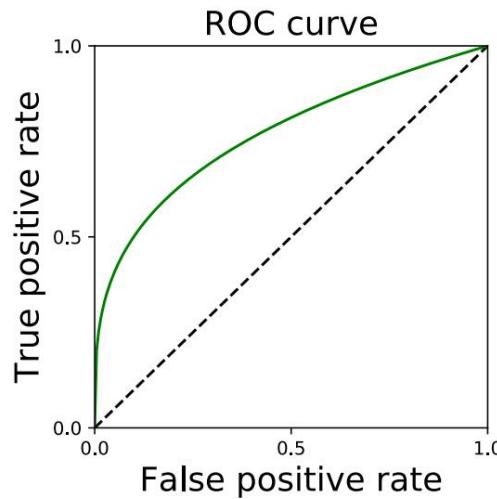


Figura 3.3: Esempio di curva ROC

Soglie variabili e curve ROC

Nel predittore ottimale per l'errore di classificazione abbiamo scelto una soglia di 1/2. Questo numero esatto era una conseguenza del costo uguale per falsi positivi e falsi negativi. Se un falso positivo fosse significativamente più costoso, potremmo voler scegliere una soglia più alta per dichiarare un positivo. Ogni scelta di una soglia comporta uno specifico compromesso tra tasso di veri positivi e tasso di falsi positivi. Variando la soglia da 0 a 1, possiamo tracciare una curva in uno spazio bidimensionale dove gli assi corrispondono al tasso di veri positivi e al tasso di falsi positivi. Questa curva è chiamata curva ROC. ROC sta per caratteristica dell'operatore del ricevitore, un nome che punta alle radici del concetto nell'elaborazione del segnale.

Nella teoria delle decisioni statistiche, la curva ROC è una proprietà di una distribuzione (X,Y). Ci fornisce per ogni impostazione del tasso di falsi positivi, il tasso ottimale di veri positivi che può essere ottenuto per il dato tasso di falsi positivi sulla distribuzione (X,Y). Ciò porta a diverse proprietà teoriche interessanti della curva ROC. Nel contesto del machine learning, le curve ROC vengono calcolate più liberamente per ogni dato punteggio di rischio, anche se non è ottimale. La curva ROC viene spesso utilizzata per osservare quanto il nostro punteggio sia predittivo rispetto alla variabile target. Una misura comune di predittività è l'area sotto la curva (AUC), che equivale alla probabilità che un'istanza casuale positiva ottenga un punteggio superiore a un'istanza casuale negativa. Un'area pari a 1/2 corrisponde a un'ipotesi casuale e un'area pari a 1 corrisponde alla classificazione perfetta.

Apprendimento supervisionato

L'apprendimento supervisionato è ciò che rende la classificazione algoritmica. Riguarda come costruire buoni classificatori a partire da campioni estratti da una popolazione. I dettagli dell'apprendimento supervisionato non avranno importanza in questo capitolo, ma ne vale comunque la pena

una comprensione operativa dell'idea di base.

Supponiamo di aver etichettato i dati, chiamati anche esempi di addestramento, della forma $(x_1, y_1), \dots, (x_n, y_n)$, dove ogni esempio è una coppia (x_i, y_i) di un'istanza x_i e un'etichetta y_i . Solitamente presupponiamo che questi esempi siano stati estratti indipendentemente e ripetutamente dalla stessa distribuzione (X, Y) . Un algoritmo di apprendimento supervisionato prende esempi di training e restituisce un classificatore, tipicamente una soglia di un punteggio: $f(x) = 1\{r(x) > t\}$. Un semplice esempio di algoritmo di apprendimento è il familiare metodo dei minimi quadrati che tenta di minimizzare la funzione obiettivo

$$\sum_{i=1}^N (r(x_i) - y_i)^2.$$

Abbiamo visto in precedenza che, a livello di popolazione, il punteggio ottimale è l'aspettativa condizionata $r(x) = E[Y | X = x]$. Il problema è che non disponiamo necessariamente di dati sufficienti per stimare ciascuna delle probabilità condizionali richieste per costruire questo punteggio. Dopotutto, il numero di possibili valori che x può assumere è esponenziale nel numero di covariate.

L'intero trucco nell'apprendimento supervisionato è approssimare questa soluzione ottimale con soluzioni algoritmamente fattibili. In tal modo, l'apprendimento supervisionato deve negoziare un equilibrio lungo tre assi:

- **Rappresentazione:** scegliere una famiglia di funzioni da cui proviene il punteggio r . Una scelta comune sono le funzioni lineari $r(x) = w^\top x$ che prendono il prodotto interno delle covariate x con un vettore di coefficienti w . Rappresentazioni più complesse coinvolgono funzioni non lineari, come le reti neurali artificiali. Questa famiglia di funzioni è spesso chiamata classe del modello e i coefficienti w sono chiamati parametri del modello.
- **Ottimizzazione:** risolvere il problema di ottimizzazione risultante trovando i parametri del modello che minimizzano la funzione di perdita negli esempi di training.
- **Generalizzazione:** garantire che una piccola perdita negli esempi di formazione implica una piccola perdita nella popolazione da cui abbiamo tratto gli esempi di formazione.

I tre obiettivi dell'apprendimento supervisionato sono intrecciati. Una rappresentazione potente potrebbe rendere più semplice l'espressione di modelli complicati, ma potrebbe anche gravare sull'ottimizzazione e sulla generalizzazione. Allo stesso modo, esistono trucchi per rendere fattibile l'ottimizzazione a scapito della rappresentazione o della generalizzazione.

Per il resto di questo capitolo, possiamo pensare all'apprendimento supervisionato come a una scatola nera che ci fornisce classificatori quando vengono forniti dati di addestramento etichettati. Ciò che conta è quali proprietà hanno questi classificatori a livello di popolazione. A livello di popolazione, interpretiamo un classificatore come una variabile casuale considerando $Y = f(X)$. Ignoriamo come Y sia stato appreso da un campione finito, quale sia la forma funzionale del classificatore e come stimiamo le varie quantità statistiche da campioni finiti. Sebbene le considerazioni sui campioni finiti siano fondamentali per l'apprendimento automatico, non sono centrali per le questioni concettuali e tecniche sull'equità di cui parleremo in questo capitolo.

Gruppi nella popolazione

Il capitolo 2 ha introdotto alcune delle ragioni per cui gli individui potrebbero voler opporsi all'uso delle regole di classificazione statistica nelle decisioni consequenziali. Passiamo ora ad una preoccupazione specifica, vale a dire la discriminazione sulla base dell'appartenenza a gruppi specifici della popolazione. La discriminazione non è un concetto generale. Si occupa di categorie socialmente rilevanti che in passato sono servite da base per trattamenti ingiustificati e sistematicamente avversi. La legge degli Stati Uniti riconosce alcune categorie protette tra cui razza, sesso (che si estende all'orientamento sessuale), religione, stato di disabilità e luogo di nascita.

In molti compiti di classificazione, le caratteristiche X codificano implicitamente o esplicitamente lo status dell'individuo in una categoria protetta. Metteremo da parte la lettera A per designare una variabile casuale discreta che cattura una o più caratteristiche sensibili. Diverse impostazioni della variabile casuale A corrispondono a diversi gruppi della popolazione reciprocamente disgiunti. La variabile casuale A è spesso chiamata attributo sensibile nella letteratura tecnica.

Si noti che formalmente possiamo sempre rappresentare un numero qualsiasi di categorie protette discrete come un singolo attributo discreto il cui supporto corrisponde a ciascuna delle possibili impostazioni degli attributi originali. Di conseguenza, la nostra trattazione formale in questo capitolo si applica al caso di più categorie protette. Questa manovra formale, tuttavia, non affronta l'importante concetto di intersezionalità che si riferisce alle forme uniche di svantaggio che i membri di più categorie protette possono sperimentare.¹⁰⁵ Il fatto che

assegniamo una variabile casuale speciale per l'appartenenza al gruppo non significa che possiamo suddividere in modo pulito l'insieme di funzionalità in due categorie indipendenti come "neutro" e "sensibile". In effetti, vedremo tra breve che un numero sufficiente di caratteristiche apparentemente neutre possono spesso fornire previsioni altamente accurate sull'appartenenza a un gruppo. Ciò non dovrebbe sorprendere. Dopotutto, se consideriamo A come la variabile obiettivo in un problema di classificazione, c'è motivo di credere che le restanti caratteristiche fornirebbero un classificatore non banale per A.

La scelta degli attributi sensibili avrà generalmente profonde conseguenze poiché decide quali gruppi della popolazione evidenziare e quali conclusioni trarre dalla nostra indagine. La tassonomia indotta dalla discretizzazione può essere di per sé fonte di danno se è troppo grossolana, troppo granulare, fuorviante o imprecisa. L'atto di classificare lo status in categorie protette e di raccogliere i dati associati può di per sé essere problematico. Riprenderemo questa importante discussione nel prossimo capitolo.

Nessuna equità per inconsapevolezza

Alcuni speravano che rimuovere o ignorare gli attributi sensibili avrebbe in qualche modo garantito l'imparzialità del classificatore risultante. Sfortunatamente, questa pratica può essere inefficace e persino dannosa.

In un set di dati tipico sono presenti molte funzionalità leggermente correlate con l'attributo sensibile. Visitando il sito pinterest.com negli Stati Uniti,

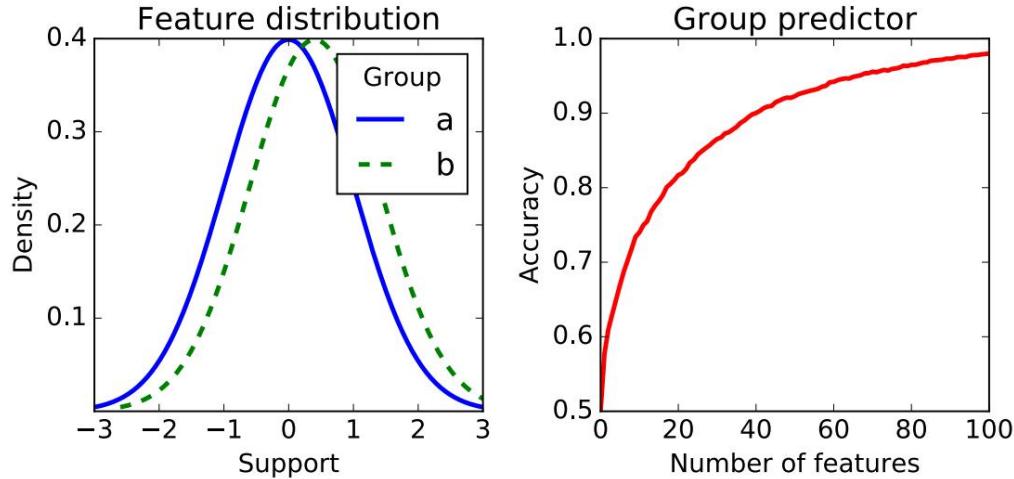


Figura 3.4: A sinistra, vediamo la distribuzione di una singola caratteristica che differisce solo leggermente tra i due gruppi. In entrambi i gruppi la caratteristica segue una distribuzione normale. Solo le medie sono leggermente diverse in ciascun gruppo. È possibile utilizzare più funzionalità come questa per creare un classificatore di appartenenza a gruppi ad alta precisione. A destra, vediamo come la precisione aumenta man mano che diventano disponibili sempre più funzionalità.

ad esempio, al momento della stesura di questo articolo aveva una piccola correlazione statistica con l'essere donna. La correlazione da sola è troppo piccola per classificare il genere di qualcuno con elevata precisione. Tuttavia, se sono disponibili numerose funzionalità di questo tipo, come nel caso di una tipica cronologia di navigazione, il compito di classificare il genere diventa fattibile a livelli di precisione più elevati.

Diverse funzionalità leggermente predittive dell'attributo sensibile possono essere utilizzate per creare classificatori ad alta precisione per tale attributo. Negli spazi di funzionalità di grandi dimensioni gli attributi sensibili sono generalmente ridondanti rispetto alle altre funzionalità. Se un classificatore addestrato sui dati originali utilizza l'attributo sensibile e rimuoviamo l'attributo, il classificatore troverà una codifica ridondante in termini di altre funzionalità. Ciò si traduce in un classificatore essenzialmente equivalente, nel senso di implementare la stessa funzione.

Per illustrare ulteriormente il problema, consideriamo una start-up fittizia che si propone di prevedere il vostro reddito in base al vostro genoma. Inizialmente, questo compito potrebbe sembrare impossibile. Come potrebbe il DNA di qualcuno rivelare il suo reddito? Tuttavia, sappiamo che il DNA codifica le informazioni sugli antenati, che a loro volta sono correlati al reddito in alcuni paesi come gli Stati Uniti. Pertanto, il DNA può probabilmente essere utilizzato per prevedere il reddito meglio delle ipotesi casuali. Il classificatore risultante utilizza l'ascendenza in modo del tutto implicito. Rimuovere le codifiche ridondanti degli antenati dal genoma è un compito difficile che non può essere portato a termine rimuovendo alcuni marcatori genetici individuali. Ciò che impariamo da questo è che l'apprendimento automatico può finire per costruire classificatori per attributi sensibili senza che venga esplicitamente richiesto, semplicemente perché è un percorso disponibile per migliorare la precisione.

Le codifiche ridondanti in genere abbondano in spazi di funzionalità di grandi dimensioni. Ad esempio, il genere può essere previsto dalle fotografie della retina con una precisione molto elevata.¹⁰⁶ E che dire dei piccoli spazi caratteristici curati manualmente? In alcuni studi, le caratteristiche vengono scelte attentamente in modo da essere più o meno statisticamente indipendenti l'una dall'altra. In questi casi, l'attributo sensibile potrebbe non avere buone codifiche ridondanti. Ciò non significa che rimuoverlo sia una buona idea. I farmaci, ad esempio, a volte dipendono legittimamente dalla razza se questi sono correlati a fattori causali sottostanti.²¹ Forzare la non correlazione dei farmaci con la razza in questi casi può danneggiare l'individuo.

Criteri statistici di non discriminazione

I criteri statistici di non discriminazione mirano a definire l'assenza di discriminazione in termini di espressioni statistiche che coinvolgono variabili casuali che descrivono una classificazione o uno scenario decisionale.

Formalmente, i criteri statistici di non discriminazione sono proprietà della distribuzione congiunta dell'attributo sensibile A, della variabile target Y, del classificatore Y o del punteggio R, e in alcuni casi anche delle caratteristiche X. Ciò significa che possiamo decidere inequivocabilmente se o nessun criterio è soddisfatto osservando la distribuzione congiunta di queste variabili casuali.

In generale, diversi criteri di equità statistica equiparano tutti alcune quantità statistiche dipendenti dal gruppo tra i gruppi definiti dalle diverse impostazioni di A. Ad esempio, potremmo chiedere di uniformare i tassi di accettazione tra tutti i gruppi. Ciò corrisponde a imporre il vincolo per tutti i gruppi a e b:

$$P\{Y = 1 | A = a\} = P\{Y = 1 | A = b\} .$$

Nel caso in cui Y ∈ {0, 1} sia un classificatore binario e abbiamo due gruppi a e b, possiamo determinare se i tassi di accettazione sono uguali in entrambi i gruppi conoscendo le tre probabilità $P\{Y = 1, A = a\}$, $P\{Y = 1, A = b\}$ e $P\{A = a\}$ che specificano completamente la distribuzione congiunta di Y e A. Possiamo anche stimare le probabilità rilevanti dati campioni casuali dalla distribuzione congiunta utilizzando argomenti statistici standard che sono non è il focus di questo capitolo.

I ricercatori hanno proposto dozzine di criteri diversi, ciascuno cercando di catturare intuizioni diverse su ciò che è giusto. Semplificando il panorama dei criteri di equità, possiamo dire che ce ne sono essenzialmente tre fondamentalmente diversi. Ciascuno di questi equivale a una delle seguenti tre statistiche in tutti i gruppi:

- Tasso di accettazione $P\{Y = 1\}$ di un classificatore
- Tassi di errore $P\{Y = 0 | Y = 1\}$ e $P\{Y = 1 | Y = 0\}$ di un classificatore Y
- Frequenza del risultato dato il valore del punteggio $P\{Y = 1 | R = r\}$ di un punteggio R

I tre criteri possono essere generalizzati per valutare le funzioni utilizzando semplici dichiarazioni di indipendenza (condizionate). Usiamo la notazione $U \perp\!\!\!\perp V | W$ per denotare che le variabili casuali U e V sono condizionatamente indipendenti dato W. Ciò significa che, a condizione che venga impostata $W = w$, le variabili casuali U e V sono indipendenti.

Tabella 3.3: Criteri di non discriminazione

Sufficienza della separazione dell'indipendenza	
R $\ddot{\gamma}$ A	R $\ddot{\gamma}$ A AA $\ddot{\gamma}$ A R

Di seguito introdurremo e discuteremo ciascuna di queste condizioni in dettaglio. Questo capitolo si concentra sulle proprietà matematiche e sulle relazioni tra questi diversi criteri. Una volta acquisita familiarità con la questione tecnica, affronteremo un dibattito più ampio attorno al contenuto morale e normativo di queste definizioni nel capitolo 4.

Indipendenza

Il nostro primo criterio formale richiede che la caratteristica sensibile sia statisticamente indipendente dal punteggio.

Definizione 1. Le variabili casuali (A, R) soddisfano l'indipendenza se $A \ddot{\gamma} R$.

Se R è una funzione di punteggio che soddisfa l'indipendenza, allora qualsiasi classificatore $Y = 1\{R > t\}$ che limita il punteggio a un valore t soddisfa anch'esso l'indipendenza. Ciò è vero finché la soglia è indipendente dall'appartenenza al gruppo. Le soglie specifiche del gruppo potrebbero non preservare l'indipendenza.

L'indipendenza è stata esplorata attraverso molte definizioni equivalenti e correlate. Quando applicata a un classificatore binario Y, l'indipendenza viene spesso definita parità demografica, parità statistica, equità di gruppo, impatto disparato e altri. In questo caso l'indipendenza corrisponde alla condizione

$$P\{Y = 1 | A = a\} = P\{Y = 1 | A = b\},$$

per tutti i gruppi a, b. Considerando l'evento $Y = 1$ come "accettazione", la condizione richiede che il tasso di accettazione sia lo stesso in tutti i gruppi. Un allentamento del vincolo introduce una quantità positiva di margine di flessibilità > 0 e lo richiede

$$P\{Y = 1 | A = a\} \ddot{\gamma} P\{Y = 1 | A = b\} \ddot{\gamma}$$

Nota che possiamo scambiare a e b per ottenere una disegualanza nella direzione opposta. Un rilassamento alternativo consiste nel considerare una condizione di rapporto, come ad esempio:

$$\frac{P\{Y = 1 | A = a\} \ddot{\gamma} 1}{P\{Y = 1 | A = b\}}.$$

Alcuni hanno sostenuto che, per $= 0,2$, questa condizione si riferisce alla regola dell'80 % che appare nelle discussioni sulle leggi sugli impatti disparati.¹⁰⁷

Ancora un altro modo per affermare la condizione di indipendenza in piena generalità è richiedere che A e R debbano avere zero informazioni reciproche. $I(A; R) = 0$. L'informazione reciproca quantifica la quantità di informazioni rivelate da una variabile casuale

riguardo all'altro. Possiamo definirlo in termini della funzione entropica più standard come $I(A; R) = H(A) + H(R) - H(A, R)$. La caratterizzazione in termini di mutua informazione porta ad utili allentamenti del vincolo. Ad esempio, potremmo richiedere $I(A; R) \leq \epsilon$.

Limitazioni dell'indipendenza

L'indipendenza è perseguita come criterio in molti articoli, per molteplici ragioni. Alcuni sostengono che la condizione riflette un presupposto di uguaglianza: tutti i gruppi hanno pari diritto di accettazione e le risorse dovrebbero quindi essere allocate proporzionalmente. Ciò che incontriamo qui è una questione sul significato normativo dell'indipendenza, che approfondiremo nel capitolo 4. Ma c'è anche una ragione più banale per la prevalenza di questo criterio. L'indipendenza ha proprietà tecniche convenienti, che rendono il criterio attraente per i ricercatori di machine learning. Spesso è quello più semplice con cui lavorare matematicamente e algoritmicamente.

Tuttavia, le decisioni basate su un classificatore che soddisfa l'indipendenza possono avere proprietà indesiderabili (e argomenti simili si applicano ad altri criteri statistici).

Ecco un modo in cui ciò può accadere, che è più facile da illustrare se immaginiamo un decisore insensibile o mal intenzionato. Immaginate un'azienda che nel gruppo a assume candidati accuratamente selezionati ad un certo tasso $p > 0$. Nel gruppo b, l'azienda assume candidati selezionati con noncuranza allo stesso tasso p . Anche se i tassi di accettazione in entrambi i gruppi sono identici, è molto più probabile che i candidati non qualificati vengano selezionati in un gruppo piuttosto che nell'altro. Di conseguenza, col senso di poi sembrerà che i membri del gruppo b abbiano ottenuto risultati peggiori dei membri del gruppo a, stabilendo così un track record negativo per il gruppo b.

Un fenomeno reale simile a questo ipotetico esempio è chiamato la scogliera di vetro: le donne e le persone di colore hanno maggiori probabilità di essere nominate CEO quando un'azienda è in difficoltà. Quando l'azienda ottiene scarsi risultati durante il suo mandato, è probabile che venga sostituita da uomini

bianchi.^{108, 109} Questa situazione potrebbe verificarsi senza che vi sia malizia: l'azienda potrebbe storicamente assumere dipendenti principalmente dal gruppo A, dando loro una migliore comprensione questo gruppo. Dal punto di vista tecnico, l'azienda potrebbe avere sostanzialmente più dati di addestramento nel gruppo a, portando così potenzialmente a tassi di errore inferiori di un classificatore appreso all'interno di quel gruppo. L'ultimo punto è un po' sottile. Dopotutto, se entrambi i gruppi fossero del tutto omogenei sotto tutti gli aspetti rilevanti per il compito di classificazione, più dati di addestramento in un gruppo sarebbero ugualmente vantaggiosi per entrambi. D'altra parte, il semplice fatto che abbiamo scelto di distinguere questi due gruppi indica che ritieniamo che potrebbero essere eterogenei sotto aspetti rilevanti.

Separazione

Il nostro criterio successivo riguarda la limitazione dell'indipendenza che abbiamo descritto. In un tipico problema di classificazione, c'è differenza tra accettare un'istanza positiva o accettare un'istanza negativa. La variabile obiettivo Y suggerisce un modo per suddividere la popolazione in strati con pari pretese di accettazione. Visto questo

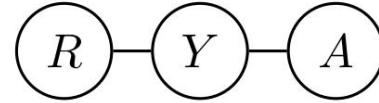


Figura 3.5: Rappresentazione del modello grafico della separazione

In questo modo, la variabile target ci dà un senso di merito. Un particolare gruppo demografico ($A = a$) può essere più o meno ben rappresentato in questi diversi strati definiti dalla variabile target. Un decisore potrebbe sostenere che in questi casi è giustificato accettare più o meno individui del gruppo a.

Queste considerazioni motivano un criterio che richiede l'indipendenza all'interno di ciascuno strato della popolazione definito dalla variabile obiettivo. Possiamo formalizzare questo requisito utilizzando una dichiarazione di indipendenza condizionale.

Definizione 2. Le variabili casuali (R, A, Y) soddisfano la separazione se $R \perp\!\!\!\perp A | Y$.

La dichiarazione di indipendenza condizionale si applica anche se le variabili assumono più di due valori ciascuna. Ad esempio, la variabile target potrebbe suddividere la popolazione in molti tipi diversi di individui.

Possiamo visualizzare la separazione come un modello grafico in cui R è separato da A dalla variabile target Y : se non hai

mai visto modelli grafici prima, non preoccuparti. Tutto ciò dice che R è condizionatamente indipendente da A dato Y .

Nel caso di un classificatore binario, la separazione equivale a richiedere per tutti i gruppi a, b i due vincoli

$$\begin{aligned} P\{Y = 1 | Y = 1, A = a\} &= P\{Y = 1 | Y = 1, A = b\} \\ P\{Y = 1 | Y = 0, A = a\} &= P\{Y = 1 | Y = 0, A = b\}. \end{aligned}$$

Ricordiamo che $P\{Y = 1 | Y = 1\}$ è chiamato il vero tasso positivo del classificatore. È la velocità con cui il classificatore riconosce correttamente le istanze positive. Il tasso di falsi positivi $P\{Y = 1 | Y = 0\}$ evidenzia la velocità con cui il classificatore assegna erroneamente risultati positivi a istanze negative. Ricordiamo che il tasso di veri positivi è uguale a 1 meno il tasso di falsi negativi. Ciò che la separazione richiede quindi è che tutti i gruppi sperimentino lo stesso tasso di falsi negativi e lo stesso tasso di falsi positivi.

Di conseguenza, la definizione richiede la parità del tasso di errore.

Questa interpretazione in termini di uguaglianza dei tassi di errore porta a rilassamenti naturali. Ad esempio, potremmo richiedere solo l'uguaglianza dei tassi di falsi negativi. Un falso negativo, intuitivamente parlano, corrisponde a un'opportunità negata in scenari in cui l'accettazione è desiderabile, come nel caso delle assunzioni. Al contrario, quando il compito è identificare individui ad alto rischio, come nel caso della previsione del default del prestito, è comune denotare il risultato indesiderato come la classe "positiva". Ciò inverte il significato di falsi positivi e falsi negativi ed è una frequente fonte di confusione terminologica.

Perché equalizzare i tassi di errore?

L'idea di livellare i tassi di errore è stata oggetto di critiche. Gran parte del dibattito ha a che fare con il fatto che un predittore ottimale non deve necessariamente avere tassi di errore uguali in tutti i gruppi. Nello specifico, quando la propensione ai risultati positivi ($P\{Y = 1\}$) differisce tra i gruppi, un predittore ottimale avrà generalmente tassi di errore diversi. In questi casi, imporre l'uguaglianza dei tassi di errore porta a un predittore che in alcuni gruppi funziona peggio di quanto potrebbe essere. In che senso è giusto?

Una risposta è che la separazione pone l'accento sulla domanda: chi sostiene il costo di un'errata classificazione? Una violazione della separazione evidenzia il fatto che gruppi diversi sperimentano costi diversi derivanti da un'errata classificazione. Si teme che tassi di errore più elevati coincidano con gruppi storicamente emarginati e svantaggiati, causando così ulteriori danni a questi gruppi.

L'atto di misurare e segnalare i tassi di errore specifici del gruppo può creare un incentivo per i decisori a lavorare per migliorare i tassi di errore attraverso la raccolta di set di dati migliori e la costruzione di modelli migliori. Se non c'è modo di migliorare i tassi di errore in alcuni gruppi rispetto ad altri, ciò solleva dubbi sull'uso legittimo dell'apprendimento automatico in questi casi. Torneremo su questa questione normativa nei capitoli successivi.

Una seconda linea di preoccupazione riguardo al criterio di separazione riguarda l'uso della variabile obiettivo come sostituto del merito. I ricercatori hanno giustamente sottolineato che in molti casi i professionisti dell'apprendimento automatico utilizzano variabili target che riflettono la diseguaglianza e l'ingiustizia esistenti. In questi casi, soddisfare la separazione rispetto a una variabile target inadeguata non va bene. Questa valida preoccupazione, tuttavia, si applica anche all'uso dell'apprendimento supervisionato in generale in questi casi. Se non riusciamo a concordare una variabile target adeguata, l'azione giusta potrebbe essere quella di sospendere l'uso dell'apprendimento supervisionato.

Queste osservazioni suggeriscono il ruolo sottile svolto dai criteri di non discriminazione. Invece di presentare vincoli per i quali possiamo ottimizzare senza ulteriori riflessioni, possono aiutare a far emergere problemi con l'uso dell'apprendimento automatico in scenari specifici.

Visualizzare la separazione

Un classificatore binario che soddisfa la separazione deve raggiungere gli stessi tassi di veri positivi e gli stessi tassi di falsi positivi in tutti i gruppi. Possiamo visualizzare questa condizione tracciando curve ROC specifiche del gruppo.

Vediamo le curve ROC di un punteggio visualizzate per ciascun gruppo separatamente. I due gruppi hanno curve diverse che indicano che non tutti i compromessi tra tasso di veri e falsi positivi sono realizzabili in entrambi i gruppi. I compromessi ottenibili in entrambi i gruppi sono proprio quelli che si trovano sotto entrambe le curve, corrispondenti all'intersezione delle regioni racchiuse dalle curve.

La regione evidenziata è la regione possibile dei compromessi che possiamo ottenere in tutti i gruppi. Tuttavia, le soglie che raggiungono questi compromessi sono in generale anche specifiche del gruppo. In altre parole, il livello di accettazione varia a seconda del gruppo. I compromessi che non si trovano esattamente sulle curve, ma piuttosto all'interno della regione, richiedono la randomizzazione. Per comprendere questo punto, pensa a come possiamo realizzare dei compromessi

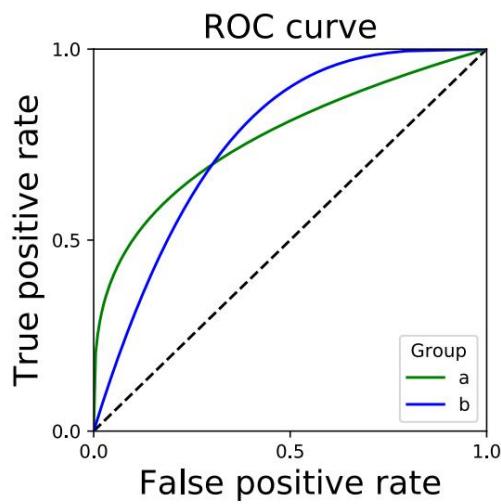


Figura 3.6: Curva ROC per gruppo.

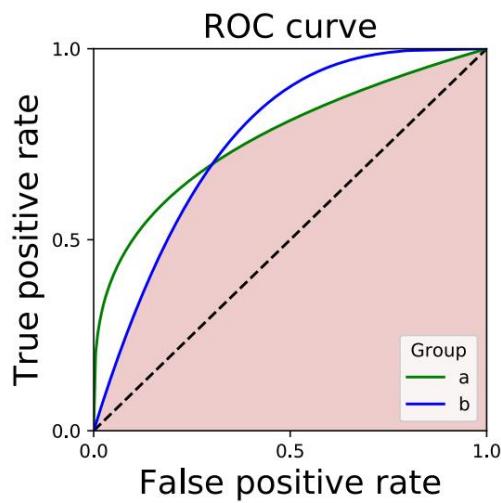


Figura 3.7: Intersezione dell'area sotto le curve.

sulla linea tratteggiata nella trama. Prendi un classificatore che accetti tutti. Ciò corrisponde al tasso di veri e falsi positivi 1, raggiungendo quindi l'angolo in alto a destra del grafico. Prendi un altro classificatore che non accetta nessuno, risultando in un tasso di veri e falsi positivi pari a 0, nell'angolo inferiore sinistro del grafico. Costruiamo ora un terzo classificatore che, data un'istanza, scelga e applichi casualmente il primo classificatore con probabilità $1 - p$ e il secondo con probabilità p . Questo classificatore raggiunge il tasso di veri e falsi positivi p dandoci così un punto sulla linea tratteggiata nel grafico. Allo stesso modo, avremmo potuto scegliere qualsiasi altra coppia di classificatori e randomizzarli tra loro. In questo modo possiamo realizzare l'intera area sotto la ROC

curva.

Tariffe di accettazione condizionate

Un parente dei criteri di indipendenza e separazione è comune nei dibattiti sulla discriminazione. Qui designiamo una variabile casuale W e chiediamo l'indipendenza condizionata della decisione Y e lo stato del gruppo A condizionato alla variabile W . Cioè, per tutti i valori w che W potrebbe assumere, e per tutti i gruppi a e b , chiediamo:

$$P\{Y = 1 | W = w, A = a\} = P\{Y = 1 | W = w, A = b\}$$

Formalmente ciò equivale a sostituire Y con W nella nostra definizione di separazione. Spesso W corrisponde a un sottoinsieme delle covariate di X . Ad esempio, potremmo richiedere che l'indipendenza valga per tutti gli individui con pari livello di istruzione. In questo caso, sceglieremmo W per riflettere il livello di istruzione raggiunto. In tal modo, diamo la licenza al decisore di distinguere tra individui con background educativi diversi. Quando applichiamo questo criterio, l'onere ricade sulla scelta corretta di cosa condizionare, il che determina se rileviamo o meno una discriminazione. In particolare, dobbiamo stare attenti a non condizionarci al meccanismo con cui il decisore discrimina. Ad esempio, un decisore mal intenzionato potrebbe discriminare imponendo requisiti educativi eccessivi per un lavoro specifico, sfruttando il fatto che questo livello di istruzione è distribuito in modo non uniforme tra i diversi gruppi. Potremo ritornare alla questione su cosa condizionare con molta più sostanza una volta che avremo acquisito familiarità con la causalità nel capitolo 5.

Sufficienza

Il nostro terzo criterio formalizza che il punteggio sussume già la caratteristica sensibile allo scopo di prevedere l'obiettivo. Anche questa idea si riduce a una dichiarazione di indipendenza condizionale.

Definizione 3. Diciamo che le variabili casuali (R, A, Y) soddisfano la sufficienza se $Y \perp\!\!\!\perp A | R$.

Possiamo visualizzare la sufficienza come modello grafico come abbiamo fatto prima con la separazione.

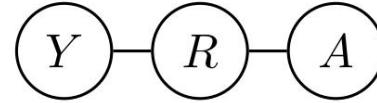


Figura 3.8: Rappresentazione del modello grafico della sufficienza

Scriviamo la definizione in modo più esplicito nel caso binario in cui $Y \in \{0, 1\}$. In questo caso, una variabile casuale R è sufficiente per A se e solo se per tutti i gruppi a, b e tutti i valori r nel supporto di R , abbiamo

$$P\{Y = 1 | R = r, A = a\} = P\{Y = 1 | R = r, A = b\}.$$

Se sostituiamo R con un predittore binario Y , riconosciamo che questa condizione richiede una parità di valori predittivi positivi/negativi in tutti i gruppi.

Calibrazione e sufficienza

La sufficienza è strettamente correlata a un concetto importante chiamato calibrazione. In alcune applicazioni è auspicabile poter interpretare i valori delle funzioni punteggio come se fossero probabilità. La nozione di calibrazione ci permette di muoverci in questa direzione. Restringendo la nostra attenzione alle variabili di risultato binarie, diciamo che un punteggio R è calibrato rispetto a una variabile di risultato Y se per tutti i valori di punteggio $r \in [0, 1]$, abbiamo

$$P\{Y = 1 | R = r\} = r.$$

Questa condizione significa che l'insieme di tutte le istanze a cui è assegnato un valore di punteggio r ha una frazione r di istanze positive tra di loro. La condizione si riferisce al gruppo di tutti gli individui che ricevono un particolare valore di punteggio. Non è necessario che la calibrazione venga effettuata nei sottogruppi della popolazione. In particolare è importante non interpretare il punteggio come una probabilità individuale. La calibrazione non ci dice nulla sul risultato di un individuo specifico che riceve un valore particolare.

Dalla definizione si vede che la sufficienza è strettamente legata all'idea di calibrazione. Per formalizzare il collegamento diciamo che il punteggio R soddisfa la calibrazione per gruppo se soddisfa

$$P\{Y = 1 | R = r, A = a\} = r,$$

per tutti i valori di punteggio r e i gruppi a . Si osservi che la calibrazione è lo stesso requisito a livello di popolazione senza il condizionamento su A .

Fatto 2. La calibrazione per gruppo implica la sufficienza.

Al contrario, la sufficienza è solo leggermente più debole della calibrazione per gruppo, nel senso che una semplice ridenominazione dei valori del punteggio passa da una proprietà all'altra.

Proposizione 1. Se un punteggio R soddisfa la sufficienza, allora esiste una funzione $: [0, 1] \rightarrow [0, 1]$ tale che (R) soddisfa la calibrazione per gruppo.

Prova. Fissare qualsiasi gruppo a e porre $(r) = P\{Y = 1 | R = r, A = a\}$. Poiché R soddisfa la sufficienza, questa probabilità è la stessa per tutti i gruppi a e quindi questa mappa è la stessa indipendentemente dal valore a che abbiamo scelto.

Consideriamo ora due gruppi qualsiasi a, b. Abbiamo,

$$\begin{aligned} r &= P\{Y = 1 | (R) = r, A = a\} \\ &= P\{Y = 1 | R \ddot{\gamma}^{\circ 1} (r), A = a\} \\ &= P\{Y = 1 | R \ddot{\gamma}^{\circ 1} (r), A = b\} \\ &= P\{Y = 1 | (R) = r, A = b\}, \end{aligned}$$

mostrando così che (R) è calibrato per gruppo. □

Concludiamo che la sufficienza e la calibrazione per gruppo sono nozioni essenzialmente equivalenti.

In pratica, esistono varie euristiche per ottenere la calibrazione. Ad esempio, il ridimensionamento di Platt prende un punteggio possibilmente non calibrato, lo tratta come una singola caratteristica e adatta un modello di regressione a una variabile rispetto alla variabile target basata su questa caratteristica.¹¹⁰ Applichiamo anche il ridimensionamento di Platt per ciascuno dei gruppi definiti dall'attributo sensibile .

Calibrazione per gruppo come conseguenza dell'apprendimento non vincolato

La sufficienza è spesso soddisfatta dal risultato di un apprendimento supervisionato e non vincolato senza la necessità di alcun intervento esplicito. Ciò non dovrebbe sorprendere. Dopotutto, l'obiettivo dell'apprendimento supervisionato è approssimare una funzione di punteggio ottimale. La funzione di punteggio ottimale che abbiamo visto in precedenza, tuttavia, è calibrata per qualsiasi gruppo, come afferma formalmente il fatto successivo.

Fatto 3. Il punteggio ottimale $r(x) = E[Y | X = x]$ soddisfa la calibrazione di gruppo per qualsiasi gruppo.

Nello specifico, per ogni insieme S che abbiamo

$$P\{Y = 1 | R = r, X \ddot{\gamma} S\} = r.$$

Generalmente ci aspettiamo che un punteggio appreso soddisfi la sufficienza nei casi in cui l'appartenenza al gruppo è esplicitamente codificata nei dati o può essere prevista da altri attributi. Per illustrare questo punto, esaminiamo i valori di calibrazione di un modello standard di apprendimento automatico, un insieme di foreste casuali, in un compito di classificazione del reddito derivato dall'American Community Survey dell'US Census Bureau.¹¹¹ Limitiamo il set di dati ai tre stati più popolosi , California, Texas e Florida.

Dopo aver suddiviso i dati in dati di training e di test, adattiamo un insieme di foreste casuali utilizzando la libreria Python standard sklearn sui dati di training. Esaminiamo quindi quanto il modello sia ben calibrato e pronto all'uso sui dati di test.

Vediamo che le curve di calibrazione per i tre gruppi razziali più grandi nel set di dati, che il Census Bureau codifica come "solo bianchi", "solo neri o afroamericani" e "solo asiatici", sono molto vicine alla diagonale principale. Questo

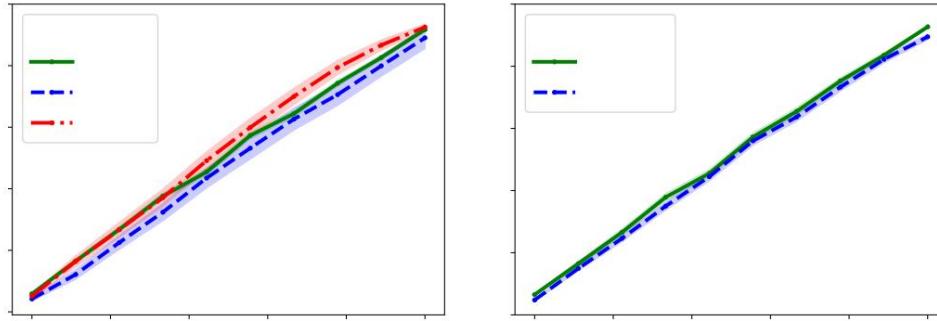


Figura 3.9: Curve di calibrazione di gruppo sui dati Census ACS

significa che i punteggi derivati dal nostro modello di foresta casuale soddisfano la calibrazione per gruppo fino a un piccolo errore. Lo stesso vale se si considerano i due gruppi "Maschile" e "Femminile" nel set di dati.

Queste osservazioni non sono una coincidenza. La teoria mostra che, in determinate condizioni tecniche, l'apprendimento supervisionato e non vincolato implica, di fatto, una calibrazione del gruppo.¹¹² Si noti, tuttavia, che affinché ciò sia vero, il classificatore deve essere in grado di rilevare l'appartenenza al gruppo. Se il rilevamento dell'appartenenza al gruppo è impossibile, la calibrazione del gruppo generalmente fallisce.

La lezione è che la sufficienza spesso arriva gratuitamente (almeno approssimativamente) come conseguenza delle pratiche standard di apprendimento automatico. Il rovescio della medaglia è che imporre la sufficienza come vincolo su un sistema di classificazione potrebbe non essere un grande intervento. In particolare, non comporterebbe un cambiamento sostanziale nelle pratiche attuali.

Come soddisfare un criterio di non discriminazione

Ora che abbiamo formalmente introdotto tre criteri di non discriminazione, vale la pena chiederci come possiamo realizzarli algoritmicamente. Distinguiamo tre diverse tecniche. Sebbene generalmente si applichino a tutti i criteri e alle relative attenuazioni esaminati in questo capitolo, la nostra discussione qui si concentra sull'indipendenza.

- **Pre-elaborazione:** regola lo spazio delle funzionalità in modo che non sia correlato con gli aspetti sensibili attributo.
- **In formazione:** inserire il vincolo nel processo di ottimizzazione che costruisce un classificatore dai dati di addestramento.
- **Post-elaborazione:** modificare un classificatore appreso in modo che non sia correlato all'attributo sensibile.

I tre approcci hanno diversi punti di forza e di debolezza.

La pre-elaborazione è una famiglia di tecniche per trasformare uno spazio caratteristico in una rappresentazione che nel suo insieme è indipendente dall'attributo sensibile. Questo approccio è generalmente indipendente da ciò che facciamo con il nuovo spazio di funzionalità nelle applicazioni downstream. Dopo che la trasformazione pre-elaborazione garantisce l'indipendenza, anche qualsiasi processo di addestramento deterministico sul nuovo spazio soddisferà l'indipendenza. Questa è una conseguenza formale della ben nota disuguaglianza nell'elaborazione dei dati derivante dalla teoria

dell'informazione.¹¹³ Raggiungere l'indipendenza al momento dell'addestramento può portare alla massima utilità poiché riusciamo a ottimizzare il classificatore con questo criterio in mente. Lo svantaggio è che abbiamo bisogno di accedere ai dati grezzi e alla pipeline di formazione. Rinunciamo anche a un po' di generalità poiché questo approccio si applica tipicamente a specifiche classi di modelli o problemi di ottimizzazione.

La post-elaborazione si riferisce al processo di prendere un classificatore addestrato e di adattarlo possibilmente in base all'attributo sensibile e alla casualità aggiuntiva in modo tale da ottenere l'indipendenza. Formalmente, diciamo che un classificatore derivato $Y = F(R, A)$ è una funzione possibilmente randomizzata di un dato punteggio R e dell'attributo sensibile. Dato un costo per falsi negativi e falsi positivi, possiamo trovare il classificatore derivato che minimizza il costo atteso di falsi positivi e falsi negativi soggetto al vincolo di equità in questione. La post-elaborazione ha il vantaggio di funzionare con qualsiasi classificatore a scatola nera, indipendentemente dal suo funzionamento interno. Non è necessaria una riqualificazione, il che è utile nei casi in cui il percorso di formazione è complesso. Spesso è anche l'unica opzione disponibile quando abbiamo accesso solo a un modello addestrato senza alcun controllo sul processo di formazione.

La post-elaborazione a volte arriva anche con una garanzia di ottimalità: se post-elaboriamo il punteggio ottimale di Bayes per ottenere la separazione, allora il classificatore risultante sarà ottimale tra tutti i classificatori che soddisfano la separazione.¹¹⁴ La saggezza convenzionale vuole che alcuni modelli di machine learning, come il gradiente gli alberi decisionali potenziati, sono spesso quasi ottimali per Bayes su set di dati tabulari con molte più righe che colonne. In questi casi, la post-elaborazione mediante la regolazione delle soglie è quasi ottimale.

Un'obiezione comune alla post-elaborazione, tuttavia, è che il classificatore risultante utilizza l'appartenenza al gruppo in modo abbastanza esplicito impostando soglie di accettazione diverse per gruppi diversi.

Relazioni tra criteri

I criteri che abbiamo esaminato vincolano la distribuzione congiunta in modi non banali. Dovremmo quindi sospettare che imponerne due qualsiasi simultaneamente vincoli eccessivamente lo spazio al punto in cui rimangono solo soluzioni degenerate. Vedremo ora che questa intuizione è in gran parte corretta. Ciò che questo dimostra, in particolare, è che se osserviamo che un criterio è valido, ci aspettiamo che gli altri vengano violati.

Indipendenza contro sufficienza

Cominciamo con una semplice proposizione che mostra come in generale indipendenza e sufficienza si escludono a vicenda. L'unico presupposto necessario qui è quello

l'attributo sensibile A e la variabile target Y non sono indipendenti. Questo è un modo diverso per dire che l'appartenenza al gruppo ha un effetto sulle statistiche della variabile target. Nel caso binario, ciò significa che un gruppo ha un tasso di risultati positivi più elevato rispetto a un altro. Consideratelo come il caso tipico.

Proposizione 2. Supponiamo che A e Y non siano indipendenti. Allora la sufficienza e l'indipendenza non possono valere entrambe.

Prova. Con la regola di contrazione per l'indipendenza condizionale,

$$A \perp\!\!\!\perp R \text{ e } A \perp\!\!\!\perp Y | R \Rightarrow A \perp\!\!\!\perp (Y, R) \Rightarrow A \perp\!\!\!\perp Y$$

Per essere chiari, $A \perp\!\!\!\perp (Y, R)$ significa che A è indipendente dalla coppia di variabili casuali (Y, R). L'eliminazione di R non può introdurre una dipendenza tra A e Y.

Al contrario,

$$A \perp\!\!\!\perp Y \Rightarrow A \perp\!\!\!\perp R \text{ oppure } A \perp\!\!\!\perp Y | UN$$

□

Indipendenza contro separazione

Un risultato analogo di mutua esclusione vale per l'indipendenza e la separazione. L'affermazione in questo caso è un po' più artificiosa e richiede il presupposto aggiuntivo che la variabile di destinazione Y sia binaria. Inoltre è necessario che il punteggio non sia indipendente dall'obiettivo. Si tratta di un presupposto piuttosto blando, dal momento che qualsiasi funzione di punteggio utile dovrebbe avere una correlazione con la variabile target.

Proposizione 3. Supponiamo che Y sia binario, A non sia indipendente da Y e R non sia indipendente da Y. Allora, indipendenza e separazione non possono valere entrambe.

Prova. Supponiamo $Y \in \{0, 1\}$. Nella sua forma contrapositiva, l'affermazione che dobbiamo mostrare è

$$A \perp\!\!\!\perp R \text{ e } A \perp\!\!\!\perp R | Y \Rightarrow A \perp\!\!\!\perp Y \text{ oppure } R \perp\!\!\!\perp Y$$

Per la legge della probabilità totale,

$$P\{R = r | A = a\} = \sum_{y \in \{0, 1\}} P\{R = r | A = a, Y = y\}P\{Y = y | A = a\}$$

Applicando l'ipotesi $A \perp\!\!\!\perp R$ e $A \perp\!\!\!\perp R | Y$, questa equazione si semplifica in

$$P\{R = r\} = \sum_{y \in \{0, 1\}} P\{R = r | Y = y\}P\{Y = y | A = a\}$$

Applicata diversamente, anche la legge della probabilità totale dà

$$P\{R = r\} = \sum_{y \in \{0, 1\}} P\{R = r | Y = y\}P\{Y = y\}$$

Combinando questo con l'equazione precedente, abbiamo

$$\ddot{\mathbf{Y}} \quad P\{R = r | Y = y\}P\{Y = y\} = \ddot{\mathbf{y}} \quad \text{si} \quad P\{R = r | Y = y\}P\{Y = y | A = a\} \quad \text{si}$$

Un esame attento rivela che quando y varia solo tra due valori, questa equazione può essere soddisfatta solo se $A \perp\!\!\!\perp Y$ o $R \perp\!\!\!\perp Y$.

Possiamo infatti riscrivere l'equazione in modo più compatto utilizzando i simboli $p = P\{Y = 0\}$, $pa = P\{Y = 0 | A = a\}$, $ry = P\{R = r | Y = y\}$, come:

$$pr_0 + (1 - p)r_1 = par_0 + (1 - pa)r_1.$$

Equivalentemente, $p(r_0 \perp\!\!\!\perp r_1) = pa(r_0 \perp\!\!\!\perp r_1)$.

Questa equazione può essere soddisfatta solo se $r_0 = r_1$, nel qual caso $R \perp\!\!\!\perp Y$, o se $p = pa$ per ogni a , nel qual caso $Y \perp\!\!\!\perp A$.

□

L'affermazione non è vera quando la variabile target può assumere più di due valori, che è un caso naturale da considerare.

Separazione versus sufficienza

Infine, ci rivolgiamo al rapporto tra separazione e sufficienza. Entrambi richiedono una relazione di indipendenza condizionale non banale tra le tre variabili A , R , Y . Imporre entrambi contemporaneamente porta ad uno spazio delle soluzioni degenere, come conferma la nostra prossima proposizione.

Proposizione 4. Assumiamo che tutti gli eventi nella distribuzione congiunta di (A, R, Y) abbiano probabilità positiva e assumiamo $A \perp\!\!\!\perp Y$. Allora, separazione e sufficienza non possono valere entrambe.

Prova. Un fatto standard (Teorema 17.2 nel testo di Wasserman115) sull'indipendenza condizionale mostra

$$A \perp\!\!\!\perp R | Y \text{ e } A \perp\!\!\!\perp Y | R \Rightarrow A \perp\!\!\!\perp (R, Y).$$

Inoltre,

$$A \perp\!\!\!\perp (R, Y) \Rightarrow A \perp\!\!\!\perp R \text{ e } A \perp\!\!\!\perp Y$$

Prendendo il contropositivo si completa la dimostrazione.

□

Per un obiettivo binario, l'ipotesi di non degenerazione nella proposizione precedente afferma che in tutti i gruppi, a tutti i valori di punteggio, abbiamo istanze sia positive che negative. In altre parole, il valore del punteggio non risolve mai completamente l'incertezza riguardo al risultato. Ricordiamo che la sufficienza vale per la funzione di punteggio ottimo di Bayes. La proposta stabilisce quindi un fatto importante: i punteggi ottimali generalmente violano la separazione.

La proposizione si applica anche ai classificatori binari. Qui, l'ipotesi dice che all'interno di ciascun gruppo il classificatore deve avere veri positivi, falsi positivi,

tassi di veri negativi e falsi negativi. Possiamo indebolire leggermente questa ipotesi e richiedere solo che il classificatore sia imperfetto, nel senso che faccia almeno una previsione falsa positiva. Ciò che è interessante nell'affermazione risultante è che la sua prova utilizza essenzialmente solo una relazione ben nota tra il tasso di vero positivo (richiamo) e il valore predittivo positivo (precisione). Questo compromesso è spesso chiamato compromesso tra richiamo di precisione.

Proposizione 5. Supponiamo che Y non sia indipendente da A e supponiamo che Y sia un classificatore binario con un tasso di falsi positivi diverso da zero. Allora la separazione e la sufficienza non possono valere entrambe.

Prova. Poiché Y non è indipendente da A devono esserci due gruppi, chiamiamoli 0 e 1, in questo modo

$$p_0 = P\{Y = 1 \mid A = 0\} = P\{Y = 1 \mid A = 1\} = p_1 .$$

Supponiamo ora che valga la separazione. Poiché il classificatore è imperfetto, ciò significa che tutti i gruppi hanno lo stesso tasso di falsi positivi diverso da zero $FPR > 0$ e lo stesso tasso di veri positivi $TPR \neq 0$. Mostreremo che la sufficienza non è valida.

Ricordiamo che nel caso binario, la sufficienza implica che tutti i gruppi abbiano lo stesso valore predittivo positivo. Il valore predittivo positivo nel gruppo a, indicato con $PPVa$, soddisfa

$$PPVa = \frac{TPR_{pa}}{TPR_{pa} + FPR(1 - pa)} .$$

Dall'espressione si vede che $PPVa = PPVb$ solo se $TPR = 0$ oppure $FPR = 0$.

Quest'ultima ipotesi è esclusa. Quindi deve essere che $TPR = 0$. Tuttavia, in questo caso, possiamo verificare che il valore predittivo negativo NPV_0 nel gruppo 0 deve essere diverso dal valore predittivo negativo NPV_1 nel gruppo 1. Ciò segue dall'espressione

$$NPVa = \frac{(1 - FPR)(1 - pa)}{(1 - TPR)pa + (1 - FPR)(1 - pa)} .$$

Quindi la sufficienza non regge. □

Nella proposizione appena dimostrata, la separazione e la sufficienza si riferiscono entrambe al classificatore binario Y . La proposizione non si applica al caso in cui la separazione si riferisce a un classificatore binario $Y = 1\{R > t\}$ e la sufficienza si riferisce alla funzione di punteggio sottostante R .

Caso di studio: punteggio del credito

Applichiamo ora alcune delle nozioni viste al credit scoring. I punteggi di credito supportano le decisioni di prestito fornendo una stima del rischio che un richiedente del prestito risulti inadempiente su un prestito. I punteggi di credito sono ampiamente utilizzati negli Stati Uniti e in altri paesi per l'assegnazione del credito, dai microprestiti ai mutui jumbo. Negli Stati Uniti, ci sono tre principali agenzie di segnalazione del credito che raccolgono dati su vari prestatori. Queste agenzie sono organizzazioni a scopo di lucro e ciascuna offre punteggi di rischio in base ai dati raccolti. I punteggi FICO sono una famiglia ben nota di

punteggi proprietari sviluppati dalla società FICO e venduti dalle tre agenzie di segnalazione del credito.

La regolamentazione delle agenzie di credito negli Stati Uniti è iniziata con il Fair Credit Reporting Act, approvato per la prima volta nel 1970, che mira a promuovere l'accuratezza, l'equità e la privacy dei consumatori delle informazioni raccolte dalle agenzie segnalanti. L'Equal Credit Opportunity Act, una legge degli Stati Uniti promulgata nel 1974, rende illegale per qualsiasi creditore discriminare qualsiasi richiedente in base a razza, colore, religione, origine nazionale, sesso, stato civile o età.

Distribuzione del punteggio

La nostra analisi si basa sui dati pubblicati dalla Federal Reserve.¹¹⁶ Il set di dati fornisce statistiche aggregate dal 2003 su un punteggio di credito, informazioni demografiche (razza o etnia, genere, stato civile) e risultati (da definire a breve).

Ci concentreremo sulle statistiche congiunte di punteggio, gara e risultato, dove gli attributi della gara assumono quattro valori descritti di seguito.

Tabella 3.4: Distribuzione del punteggio di credito per etnia

Campioni di razza o etnia con punteggio ed esito	
Bianco	133.165
Nero	18.274
ispanico	14.702
asiatico	7.906
Totale	174.047

Il punteggio utilizzato nello studio si basa sul punteggio TransUnion TransRisk.

TransUnion è un'agenzia statunitense di segnalazione del credito. Il punteggio TransRisk è a sua volta basato sui punteggi FICO. La Federal Reserve ha rinormalizzato i punteggi dello studio affinché variassero da 0 a 100, dove 0 indica il merito creditizio minimo.

Le informazioni sulla razza sono state fornite dall'Amministrazione della Previdenza Sociale, basandosi quindi su valori autodichiarati. La distribuzione cumulativa di questi punteggi di credito dipende fortemente dal gruppo razziale, come rivela la figura successiva.

Variabili prestazionali e curve ROC

Come spesso accade, la variabile di risultato è un aspetto sottile di questo insieme di dati. Vale la pena sottolinearne la definizione. Poiché il modello di punteggio è proprietario, non è chiaro quale variabile target sia stata utilizzata durante il processo di formazione. Cos'è allora che il punteggio sta cercando di prevedere? In una prima reazione, potremmo dire che l'obiettivo di un punteggio di credito è prevedere un risultato di default. Tuttavia, questa non è una nozione chiaramente definita. Le inadempienze variano in base all'importo del debito recuperato e al tempo concesso per il recupero. Ogni singolo indicatore di prestazione binario è in genere una semplificazione eccessiva.

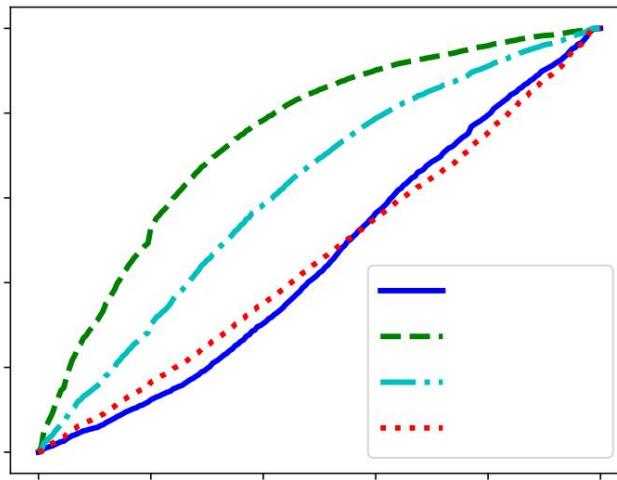


Figura 3.10: Densità cumulativa dei punteggi per gruppo.

Ciò che è disponibile nei dati della Federal Reserve è una cosiddetta variabile di performance che misura una grave inadempienza in almeno una linea di credito di un certo periodo di tempo. Più specificamente, afferma la Federal Reserve

- (la) misura si basa sulla performance di conti nuovi o esistenti e misura se gli individui sono stati in ritardo di 90 giorni o più su uno o più dei loro conti o avevano un elemento di registro pubblico o un nuovo conto dell'agenzia di riscossione durante il periodo di prestazione.

Con questa variabile di prestazione a portata di mano, possiamo osservare la curva ROC per avere un'idea di quanto sia predittivo il punteggio in diversi dati demografici.

Il significato di vero tasso positivo è il tasso di prestazione positiva prevista data la prestazione positiva. Allo stesso modo, il tasso di falsi positivi è il tasso di prestazione negativa prevista a fronte di una prestazione positiva.

Vediamo che le forme appaiono più o meno visivamente simili nei gruppi, sebbene il gruppo "Bianco" racchiuda un'area notevolmente più ampia sotto la curva rispetto al gruppo "Nero". Si noti inoltre che anche due curve ROC con la stessa forma possono corrispondere a funzioni di punteggio molto diverse. Un particolare compromesso tra tasso di veri positivi e tasso di falsi positivi raggiunto a una soglia t in un gruppo potrebbe richiedere una soglia t diversa nell'altro gruppo.

Confronto di criteri diversi

Con i dati del punteggio a portata di mano, confrontiamo quattro diverse strategie di classificazione:

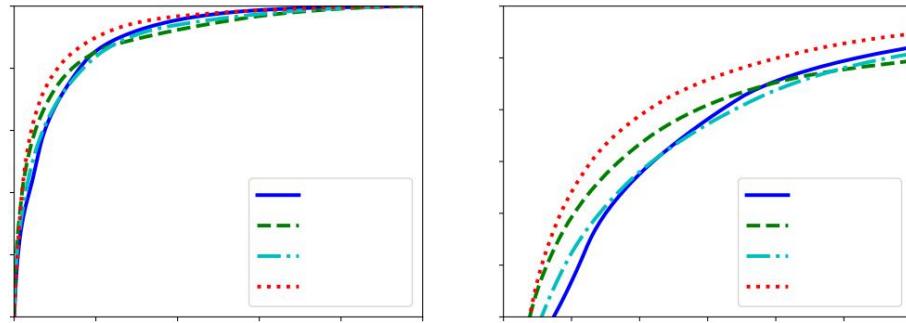


Figura 3.11: Curva ROC del punteggio di credito per gruppo.

- **Profitto massimo:** scegliere soglie di punteggio possibilmente dipendenti dal gruppo in modo da massimizzare il profitto.
- **Soglia singola:** scegli una soglia di punteggio uniforme per tutti i gruppi in modo che massimizzi il profitto.
- **Indipendenza:** raggiungere un tasso di accettazione uguale in tutti i gruppi. Soggetto a questo vincolo, massimizzare il profitto. •
- Separazione:** ottenere un tasso uguale di veri/falsi positivi in tutti i gruppi. Soggetto a questo vincolo, massimizzare il profitto.

Per dare un senso alla massimizzazione del profitto, dobbiamo assumere una ricompensa per un vero positivo (prestazione positiva prevista correttamente) e un costo per i falsi positivi (prestazione negativa prevista come positiva). Nei prestiti, il costo di un falso positivo è in genere molte volte maggiore della ricompensa per un vero positivo. In altre parole, gli interessi pagati derivanti da un prestito sono relativamente piccoli rispetto all'importo del prestito che potrebbe andare perso. A scopo illustrativo, immaginiamo che il costo di un falso positivo sia 6 volte maggiore del rendimento di un vero positivo. I numeri assoluti non contano. Conta solo il rapporto. Questa semplice struttura dei costi trascura una serie di dettagli che probabilmente sono rilevanti per il creditore, come i termini del prestito.

C'è un altro importante avvertimento sul tipo di analisi che stiamo per fare. Poiché ci vengono fornite solo statistiche aggregate, non possiamo riqualificare il punteggio tenendo presente una particolare strategia di classificazione. L'unica cosa che possiamo fare è definire un insieme di soglie che raggiungano un criterio particolare. Questo approccio può essere eccessivamente pessimistico per quanto riguarda il profitto ottenuto soggetto a ciascun vincolo. Per questo motivo e per il fatto che la nostra scelta della funzione di costo è stata piuttosto arbitraria, non indichiamo i numeri del profitto. I numeri possono essere trovati nell'analisi originale,¹¹⁴ che riporta che la "soglia unica" consente di ottenere profitti maggiori rispetto alla "separazione", che a sua volta ottiene profitti maggiori rispetto all'"indipendenza".

Ciò che facciamo invece è osservare i diversi compromessi tra vero e falso tasso positivo che ciascun criterio raggiunge in ciascun gruppo.

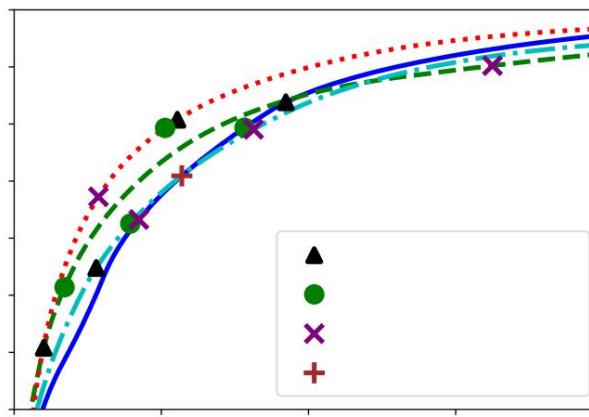


Figura 3.12: Curve ROC con soglie ottimali per diversi criteri.

Possiamo vedere che anche se le curve ROC sono in qualche modo simili, i compromessi risultanti possono differire ampiamente da gruppo a gruppo per alcuni criteri. Il vero tasso positivo ottenuto dal profitto massimo per il gruppo asiatico è il doppio di quello del gruppo nero. Il criterio di separazione, ovviamente, comporta lo stesso compromesso in tutti i gruppi. L'indipendenza eguaglia il tasso di accettazione, ma porta a compromessi molto diversi. Ad esempio, il gruppo nero ha un tasso di falsi positivi più di tre volte superiore al tasso di falsi positivi del gruppo asiatico.

Valori di calibrazione

Consideriamo infine il tasso di non default per gruppo. Ciò corrisponde al grafico di calibrazione per gruppo.

Vediamo che le curve di performance per gruppo sono ragionevolmente ben allineate. Ciò significa che una trasformazione monotona dei valori del punteggio si tradurrebbe in un punteggio approssimativamente calibrato per gruppo secondo la nostra definizione precedente. A causa delle differenze nella distribuzione del punteggio per gruppo, potrebbe tuttavia accadere che la soglia del punteggio porti a un classificatore con valori predittivi positivi diversi in ciascun gruppo. La calibrazione viene generalmente persa quando si prende un punteggio multivalore e lo si rende binario.

Limitazioni intrinseche dei criteri di osservazione

I criteri che abbiamo visto finora hanno un aspetto importante in comune. Sono proprietà della distribuzione congiunta del punteggio, dell'attributo sensibile e della variabile target. In altre parole, se conosciamo la distribuzione congiunta delle variabili casuali (R, A, Y), possiamo determinare senza ambiguità se questa distribuzione congiunta

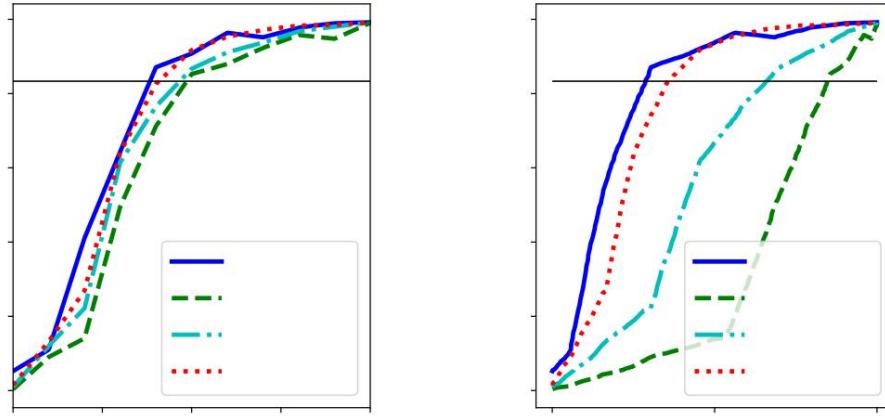


Figura 3.13: Valori di calibrazione del punteggio di credito per gruppo.

soddisfa uno di questi criteri oppure no. Ad esempio, se tutte le variabili sono binarie, ci sono otto numeri che specificano le distribuzioni congiunte. Possiamo verificare ciascuno dei criteri discussi in questo capitolo osservando solo questi otto numeri e nient'altro. Possiamo ampliare un po' questo concetto e includere anche tutte le altre funzionalità di X , non solo l'attributo group. Quindi, chiamiamo osservazionale un criterio se è una proprietà della distribuzione congiunta delle caratteristiche X , dell'attributo sensibile A , di una funzione di punteggio R e di una variabile di risultato Y . Intuitivamente parlando, un criterio è osservativo se possiamo scriverlo in modo inequivocabile utilizzando dichiarazioni di probabilità che coinvolgono le variabili casuali a portata di mano.

Le definizioni osservative hanno molti aspetti interessanti. Spesso sono facili da enunciare e richiedono solo un leggero formalismo. Non fanno alcun riferimento al funzionamento interno del classificatore, alle intenzioni del decisore, all'impatto delle decisioni sulla popolazione o ad alcuna idea su se e come una caratteristica influenzi effettivamente il risultato. Possiamo ragionare su di essi in modo abbastanza conveniente come abbiamo visto prima. In linea di principio, le definizioni osservative possono sempre essere verificate utilizzando campioni della distribuzione congiunta, soggetti a errori di campionamento statistico.

Questa semplicità delle definizioni osservative porta anche a limitazioni intrinseche. Ciò che le definizioni osservative nascondono sono i meccanismi che hanno creato una disparità osservata. In un caso, una differenza nel tasso di accettazione potrebbe essere dovuta a una considerazione dispettosa dell'appartenenza al gruppo da parte di un decisore. In un altro caso, la differenza nei tassi di accettazione potrebbe riflettere una diseguaglianza di fondo nella società che dà a un gruppo un vantaggio nell'essere accettato. Sebbene entrambi siano motivo di preoccupazione, nel primo caso la discriminazione è un'azione diretta del decisore. Nell'altro caso, il luogo della discriminazione può essere al di fuori dell'azione del decisore.

I criteri osservativi non possono, in generale, fornire risposte soddisfacenti su quali siano le cause e i meccanismi della discriminazione. I capitoli successivi, in particolare

nostro capitolo sulla causalità, sviluppare strumenti per andare oltre l'ambito dell'osservazione criteri.

Note del capitolo

Per la storia iniziale della probabilità e l'ascesa del pensiero statistico, consultare i libri di Hacking,¹¹⁷ La¹¹⁸ Porter,¹⁰³ e Desrosières.¹⁰²

teoria delle decisioni statistiche che abbiamo trattato in questo capitolo è anche chiamata teoria del rilevamento (dei segnali) ed è oggetto di vari libri di testo. Ciò che chiamiamo classificazione è chiamata anche previsione in altri contesti. Allo stesso modo, i classificatori sono spesso chiamati predittori. Per un'introduzione all'apprendimento automatico per i laureati, vedere il testo di Hardt e Recht.¹¹⁹ Il libro di testo di Wasserman¹¹⁵ fornisce un ulteriore background statistico , inclusa un'esposizione dell'indipendenza condizionale che è utile per comprendere parte del materiale del capitolo.

Criteri di equità simili a quelli esaminati in questo capitolo erano già noti negli anni '60 e '70, principalmente nella letteratura sui test educativi e sulla psicometria.¹²⁰ Il primo e più influente criterio di equità in questo contesto è dovuto a Cleary.^{121,122} Un punteggio supera il criterio di Cleary se la conoscenza dell'appartenenza al gruppo non aiuta a prevedere il risultato del punteggio con un modello lineare. Questa condizione deriva dalla sufficienza e può essere espressa sostituendo l'affermazione di indipendenza condizionale con un'affermazione analoga sulle correlazioni parziali.¹²³ Einhorn e Bass¹²⁴ hanno considerato l'uguaglianza dei valori di precisione, che è un allentamento della sufficienza come abbiamo visto

prima. Thorndike¹²⁵ ha considerato una variante debole di calibrazione in base alla quale la frequenza delle previsioni positive deve essere uguale alla frequenza dei risultati positivi in ciascun gruppo e ha proposto di ottenerla tramite una fase di post-elaborazione che fissa soglie diverse in gruppi diversi. Il criterio di Thorndike non è paragonabile alla sufficienza in generale.

Darlington¹²³ ha stabilito quattro diversi criteri in termini di espressioni succinte che coinvolgono i coefficienti di correlazione tra varie coppie di variabili casuali. Questi criteri includono l'indipendenza, l'allentamento della sufficienza, l'allentamento della separazione e il criterio di Thorndike. Darlington ha incluso un argomento visivo intuitivo che mostra che i quattro criteri sono incompatibili tranne che nei casi degenerati. Lewis¹²⁶ ha esaminato tre criteri di equità, tra cui la stessa precisione e gli stessi tassi di veri/falsi positivi.

Questi importanti primi lavori furono riscoperti più tardi nella comunità dell'apprendimento automatico e del data mining.¹²⁰ Numerosi lavori consideravano le varianti dell'indipendenza come un vincolo di equità.^{127, 128} Feldman et al.¹⁰⁷ hanno studiato un allentamento della parità demografica nel contesto di leggi sugli impatti disparati. Zemel et al.¹²⁹ hanno adottato il punto di vista dell'informazione reciproca e hanno proposto un approccio euristico di pre-elaborazione per ridurre al minimo l'informazione reciproca. Già nel 2012, Dwork et al.¹³⁰ sostenevano che il criterio di indipendenza era inadeguato come vincolo di equità. In particolare, questo lavoro ha identificato il problema dell'indipendenza di cui abbiamo discusso in questo capitolo.

Il criterio di separazione appariva sotto il nome di quote pareggiate, 114 a fianco l'allentamento dei tassi di falsi negativi, chiamato uguaglianza di opportunità. Questi criteri sono apparsi anche in un lavoro indipendente¹³¹ sotto nomi diversi. Woodworth et al.¹³² ha studiato un rilassamento della separazione espresso in termini di coefficienti di correlazione. Questo rilassamento corrisponde al terzo criterio studiato da Darlington.¹²³

ProPublica¹³³ ha implicitamente adottato l'uguaglianza dei tassi di falsi positivi come criterio di equità nel suo articolo sui punteggi COMPAS. Northpointe, il creatore del COMPAS software, hanno sottolineato l'importanza della calibrazione per gruppo nella loro confutazione¹³⁴ a L'articolo di ProPublica. Argomentazioni simili furono avanzate subito dopo la pubblicazione dell'articolo di ProPublica scritto da blogger tra cui Abe Gong. C'è stato ampio borsa di studio sulla valutazione del rischio attuariale nella giustizia penale che precede di molto il dibattito ProPublica; Berk et al.¹³⁵ forniscono un sondaggio con commenti.

Sono state mostrate varianti del compromesso tra separazione e sufficienza Chouldechova¹³⁶ e Kleinberg et al.¹³⁷ Ognuno di loro ha considerato qualcosa criteri diversi per il trade-off. L'argomentazione di Chouldechova è molto simile alla dimostrazione abbiamo presentato che invoca la relazione tra valore predittivo positivo e vero tasso positivo. Il lavoro successivo¹³⁸ considera i compromessi tra rilassato e criteri approssimativi. Gli altri risultati del trade-off presentati in questo capitolo sono nuovi a questo libro. La prova della proposizione relativa alla separazione e all'indipendenza per i classificatori binari, così come il controsenso per i classificatori ternari, è dovuto a Shira Mitchell e Jackie Shadlen, ci hanno segnalato nella comunicazione personale.

Il caso di studio sul punteggio di credito è di Hardt, Price, e Srebro¹¹⁴ Tuttavia, noi evidenziano il criterio di indipendenza nei nostri grafici, mentre gli autori dell'articolo evidenziare invece il criterio delle pari opportunità. I numeri riguardanti il la composizione razziale del set di dati proviene dalla colonna "Campione di stima" di Tabella 9 sulla pagina web del rapporto della Federal Reserve.¹¹⁶

Un dizionario di criteri

Per comodità raccogliamo di seguito alcuni criteri di equità demografica che sono stati proposti in passato (non necessariamente includendo il riferimento originale). Partiremo loro al parente più stretto tra i tre criteri: indipendenza, separazione, e sufficienza. Questa tabella è intesa solo come riferimento e non è esaustiva. Là non è necessario memorizzare questi nomi diversi.

Tabella 3.5: Elenco dei criteri statistici di non discriminazione

Nome	Nota sui criteri	Riferimento
Indipendenza	Indipendente Equiv.	Calders et al. (2009)
Equità di gruppo	Indipendente Equiv.	
Parità demografica	Indipendente Equiv.	
Parità statistica condizionale	Indipendente Relax.	Corbett-Davies et al. (2017)
Criterio Darlington (4)	Indipendente Relax.	Darlington (1971)
Pari opportunità	Separ.	Relax. Hardt, Price, Srebro (2016)
Quote pareggiate	Separ.	Equiv. Hardt, Price, Srebro (2016)

Nome	Nota sui criteri	Riferimento
Precisione condizionata della procedura Separ.	Equiv.	Berk et al. (2017)
Evitare maltrattamenti disparati Separ.	Equiv.	Zafar et al. (2017)
Saldo della classe negativa Separ.	Relax.	Kleinberg et al. (2016)
Saldo della classe positiva Separ.	Relax.	Kleinberg et al. (2016)
Uguaglianza predittiva Separ.	Relax.	Corbett-Davies et al. (2017)
Correlazioni equalizzate Separ.	Relax.	Woodworth (2017)
Criterio Darlington (3)	Separ.	Relax.
Modello Cleary Suff.	Relax.	Chiara (1966)
Precisione d'uso condizionale Suff.	Equiv.	Berk et al. (2017)
Parità predittiva Suff.	Relax.	Culdechova (2016)
Calibrazione all'interno dei gruppi Criterio Darlington (1), (2)	Suff.	Equiv.
	Suff.	Relax.
		Darlington (1971)

4

Nozioni relative di equità

Nel capitolo 3 abbiamo considerato una serie di criteri statistici che aiutano a evidenziare le differenze a livello di gruppo sia nei trattamenti che nei risultati che potrebbero essere ottenuti dall'uso di un modello di apprendimento automatico. Ma perché dovremmo preoccuparci delle differenze a livello di gruppo? E come dovremmo decidere di quali gruppi dovremmo preoccuparci? In questo capitolo esploreremo le molte diverse ragioni normative che potremmo avere per opporci alle differenze a livello di gruppo. Si tratta di un sottile, ma importante, spostamento di focus rispetto al Capitolo 2, in cui abbiamo considerato alcune delle ragioni normative per cui gli individui potrebbero opporsi a schemi decisionali che distribuiscono risorse o opportunità desiderabili. In questo capitolo ci concentreremo sul motivo per cui potremmo essere preoccupati per l'allocazione non equa di risorse e opportunità tra gruppi specifici e nella società in generale. In particolare, esamineremo i fondamenti normativi che fondono le rivendicazioni di discriminazione e le richieste di giustizia distributiva.

Cercheremo poi di collegare più direttamente queste argomentazioni ai criteri statistici sviluppati nel capitolo 3, con l'obiettivo di conferire a tali criteri maggiore sostanza normativa.

Una questione di terminologia: useremo i termini ingiustizia e discriminazione più o meno come sinonimi. Linguisticamente, il termine discriminazione pone maggiormente l'accento sull'azione del decisore. Evitiamo inoltre specificamente di utilizzare la terminologia "trattamento disparato" e "impatto disparato" in questo capitolo poiché si tratta di termini legali dell'arte con significati e significato giuridico più precisi; li affronteremo nel capitolo 6.

Svantaggio relativo sistematico

Le discussioni sulla discriminazione nel contesto dell'apprendimento automatico possono sembrare strane se si considera che lo scopo stesso di molte applicazioni di apprendimento automatico è capire come trattare persone diverse in modo diverso, ovvero discriminare tra loro. Tuttavia, ciò che chiamiamo discriminazione in questo capitolo non è un trattamento diverso in sé e per sé, ma piuttosto un trattamento che impone sistematicamente uno svantaggio a un gruppo sociale rispetto ad altri. La sistematicità nelle differenze di trattamento e di risultati è ciò che conferisce alla discriminazione la sua forza normativa come concetto.

Per comprendere meglio questo punto, consideriamo tre livelli ai quali le persone potrebbero essere soggette a un trattamento ingiusto. In primo luogo, una persona potrebbe essere soggetta al pregiudizio di

un decisore individuale, ad esempio uno specifico responsabile delle assunzioni le cui decisioni sono influenzate dall'animosità razziale. In secondo luogo, una persona potrebbe incontrare barriere sistematiche all'accesso a determinate occupazioni, forse perché i membri del gruppo a cui appartiene non sono considerati idonei a diventare ingegneri, medici, avvocati, ecc., indipendentemente dalle loro reali capacità o potenzialità. Ad esempio, in alcune occupazioni le donne potrebbero avere opportunità di lavoro limitate a causa del loro genere. Infine, alcune caratteristiche personali potrebbero costituire un principio organizzativo per la società nel suo complesso, in modo tale che i membri di determinati gruppi siano sistematicamente esclusi dalle opportunità in molteplici ambiti della vita. Ad esempio, la razza e il genere potrebbero limitare l'accesso delle persone non solo al lavoro, ma anche all'istruzione, al credito, all'alloggio, ecc.

Negli esempi del paragrafo precedente, ci siamo affidati alla razza e al genere proprio perché entrambi sono serviti, storicamente, come principi organizzativi per molte società; non sono solo le basi idiosincratiche su cui specifici datori di lavoro o professioni hanno negato ai membri di questi gruppi importanti opportunità.¹³⁹ Ciò aiuta a spiegare perché queste sono

caratteristiche di particolare preoccupazione e perché altre potrebbero non esserlo. Ad esempio, potremmo non interessarci del fatto che un particolare datore di lavoro o una particolare professione abbia sistematicamente rifiutato i candidati mancini, oltre al fatto che potremmo trovare la decisione arbitraria e quindi irrazionale, dato che la manicità potrebbe non avere nulla a che fare con la prestazione lavorativa.

Ma se la manualità diventasse la base per privare le persone di opportunità a tutti i livelli e non solo da parte di questo decisore o in questo ambito, potremmo iniziare a considerarlo problematico. Nella misura in cui la manualità determina la posizione delle persone e la posizione nella società in generale, essa assurgerebbe al livello di una caratteristica degna di particolare attenzione.¹⁴⁰

Razza e genere – tra gli altri enumerati nella legge sulla discriminazione e descritti più dettagliatamente nel capitolo 6 – assurgono a un tale livello perché sono serviti come base per perpetuare lo svantaggio relativo sistematico a tutti i livelli. In casi estremi, alcune caratteristiche possono fornire le basi per una rigida gerarchia sociale in cui i membri di diversi gruppi vengono inseriti in posizioni più o meno desiderabili nella società. Tali condizioni possono creare l'equivalente di un sistema di caste,¹⁴¹ in cui alcuni gruppi sono confinati in una posizione permanente di relativo svantaggio. È anche importante notare la minaccia unica posta dal trattamento

differenziale sulla base di caratteristiche che persistono a livello intergenerazionale. Ad esempio, si presume spesso che i bambini appartengano allo stesso gruppo razziale dei loro genitori biologici, rendendo particolarmente sistematico lo svantaggio relativo che le persone possono sperimentare a causa della loro razza: i bambini nati in famiglie che sono state ingiustamente private di risorse e opportunità avranno meno accesso a queste risorse e opportunità, limitando così fin dall'inizio della loro vita l'efficacia con cui potrebbero realizzare il proprio potenziale, anche prima di essere essi stessi soggetti a discriminazione.

¹Naturalmente, tale stratificazione non deve necessariamente avere origine da politiche formali o da una o anche solo da una piccola manciata di decisioni altamente consequenziali. Molte azioni apparentemente piccole possono rafforzare cumulativamente vantaggi e svantaggi relativi, che vanno dalla pubblicità selettiva di un lavoro attraverso il passaparola ai commenti sessisti sul posto di lavoro (maggiori informazioni su questo argomento nel Capitolo 8).

Sei resoconti dell'illegittimità della discriminazione

Gli studiosi hanno sviluppato molte teorie normative per spiegare l'illegittimità della discriminazione, in particolare l'illegittimità di trattare le persone in modo diverso in base a caratteristiche come razza, sesso o disabilità. Sebbene ciascuna di queste teorie si occupi del modo in cui tale trattamento differenziale dà origine a uno svantaggio relativo sistematico, differiscono nel modo in cui comprendono ciò che rende moralmente discutibile il processo decisionale sulla base di queste caratteristiche.

Rilevanza: uno dei motivi – e forse il motivo più comune – per opporsi alla discriminazione è perché si basa su caratteristiche che hanno poca o nessuna rilevanza per il risultato o la qualità che i decisori potrebbero cercare di prevedere o valutare.^{142, 141} Ad esempio, uno dei motivi per cui potrebbe essere sbagliato basare le decisioni relative all'assunzione su caratteristiche come la razza o il genere è che queste caratteristiche non hanno alcuna rilevanza per le determinazioni sulla capacità dei candidati di svolgere il lavoro.

Si noti che questa è una variante dell'obiezione trattata nel capitolo 2, in cui gli individui potrebbero contestare le decisioni sulla base del fatto che sono state rese sulla base di informazioni irrilevanti. In questo caso, è importante non solo che il fare affidamento su fattori irrilevanti porti a errori, ma che tali errori portino a uno svantaggio relativo sistematico per particolari gruppi sociali.

Generalizzazioni: oppure potremmo sostenere che il danno risiede nei raggruppamenti inutilmente grossolani perpetrati da decisioni prese sulla base della razza o del genere, anche se si può dimostrare che questi possiedono una certa rilevanza statistica per la decisione in questione.⁸² Ciò ci riporta a un'altra idea contenuta nel capitolo 2: che le persone meritano di essere trattate come individui e valutate in base ai loro meriti unici. Come ricorderete, l'idea intuitiva di una valutazione perfettamente individualizzata è irraggiungibile. Qualsiasi forma di giudizio basata sulle caratteristiche individuali deve basarsi su alcune generalizzazioni e sull'esperienza passata. Tuttavia potremmo ancora opporci alla grossolanità delle generalizzazioni, soprattutto se ci sono ovviamente informazioni aggiuntive che potrebbero fornire un modo più granulare – e quindi più accurato – per tracciare distinzioni. Ad esempio, potremmo obiettare se le donne fossero escluse dall'incarico di vigili del fuoco sulla base del presupposto che le donne come gruppo hanno meno probabilità di soddisfare i requisiti di forza, invece di sottoporre un test di idoneità ai candidati di tutti i sessi.

Pregiudizio: un altro argomento comune per cui la discriminazione è illecita è che equivale a una forma di processo decisionale pregiudizievole, in cui si presume che i membri di determinati gruppi abbiano uno status inferiore. Piuttosto che essere semplicemente un problema di rilevanza o granularità, come nelle prospettive precedenti discusse in questa sezione, è un problema di convinzioni, in cui i decisorи tengono interi gruppi in minore considerazione rispetto ad altri. Ad esempio, il problema con le decisioni guidate da animus razziale o misoginia non è semplicemente che possono dar luogo a previsioni imprecise, ma che i decisorи mantengono queste opinioni in primo luogo.^{143,}

^{144, 139} **Mancanza di rispetto:** un'idea correlata, ma distinta. È che prendere decisioni sulla base di tali caratteristiche è sbagliato quando sminuisce coloro che possiedono tali caratteristiche.^{139, 145} In questo senso, il problema della discriminazione è che considera certi gruppi categoricamente inferiori ad altri e quindi non degni di uguale rispetto. Questa obiezione differisce da quelle basate sul pregiudizio perché la

il danno non è localizzato nella convinzione dei decisori circa l'inferiorità dei membri di particolari gruppi, ma in ciò che le azioni dei decisori comunicano sullo status sociale dei gruppi. Ad esempio, il problema delle pratiche di assunzione sessiste non è semplicemente che confinano le donne a ruoli particolari nella società o che si basano su convinzioni pregiudiziali, ma suggeriscono che le donne sono inferiori agli uomini. Intesa in questo modo, la discriminazione è dannosa non solo per la specifica persona oggetto di una decisione sfavorevole, ma per l'intero gruppo a cui la persona appartiene perché nuoce alla posizione sociale del gruppo nella comunità.

Immutabilità: un argomento completamente diverso sul motivo per cui la discriminazione è illecita è perché implica trattare le persone in modo diverso a seconda delle caratteristiche sulle quali non hanno alcun controllo. Per questo motivo, il motivo per cui dovremmo preoccuparci delle differenze nel trattamento, ad esempio, delle persone con o senza disabilità è perché le persone con disabilità potrebbero non avere alcun controllo sulla propria disabilità.^{146, 147, 148} Come esplorato in Capitolo 2, le decisioni che si basano su caratteristiche immutabili negano alle persone che possiedono queste caratteristiche la possibilità di ottenere risultati diversi dal processo decisionale, condannando di fatto tutte queste persone a esiti negativi.^[2] Ciò equivale a un'ingiusta discriminazione, in particolare quando la caratteristica immutabile in La domanda è quella il cui utilizzo nel processo decisionale darà luogo a uno svantaggio relativo sistematico.

Aggravamento dell'ingiustizia: o forse il problema della discriminazione deriva dal fatto che essa può aggravare l'ingiustizia esistente.¹⁴⁹ In molti modi, questa è un'estensione dell'argomentazione precedente sul controllo, ma con un focus specifico sugli effetti dell'ingiustizia passata. In particolare, si sostiene che le persone non possono essere moralmente colpevoli per determinati fatti su se stessi che non sono il risultato delle proprie azioni, soprattutto se questi fatti sono il risultato di essere stati soggetti a qualche ingiustizia passata. Non tenere conto del fatto che i membri di alcuni gruppi potrebbero essere stati, in passato, soggetti a molti dei tipi di problemi sopra descritti potrebbe portare i decisorи a sentirsi perfettamente giustificati nel trattare i membri di questi gruppi in modo diverso. Tuttavia, il motivo per cui le persone potrebbero apparire diversamente al momento della decisione potrebbe essere qualche ingiustizia passata, inclusa la discriminazione p-

Ignorare questo fatto significherebbe che i decisorи sottoporrebbero gruppi specifici a decisioni peggiori semplicemente perché hanno già subito qualche danno in precedenza. Si noti che questa obiezione non ha nulla a che fare con le preoccupazioni relative all'accuratezza; infatti, suggerisce, come abbiamo discusso per la prima volta nel capitolo 2, che potremmo avere l'obbligo morale di sottovalutare talvolta gli effetti di fattori sui quali le persone non hanno alcun controllo, anche se ciò significa fare previsioni meno accurate. Nel capitolo 2 abbiamo considerato questo principio senza alcuna preoccupazione particolare per i risultati distributivi; in questo caso, possiamo adattare il principio affinché renda conto dell'illegittimità della discriminazione, sottolineando che aver subito qualche maltrattamento passato a causa dell'appartenenza di qualcuno

2Sebbene ciò sia particolarmente problematico se le stesse decisioni potessero essere prese almeno altrettanto accuratamente facendo affidamento su criteri sui quali le persone esercitano un certo controllo, potremmo comunque opporci a decisioni basate su caratteristiche immutabili anche se non ci sono mezzi alternativi per prendere decisioni a livello un pari livello di precisione. Ad esempio, potrebbe darsi che le informazioni genetiche siano la base più affidabile su cui fare previsioni sulla salute futura di una persona. Potremmo tuttavia opporci all'idea che alle persone debbano essere addebitati premi assicurativi più alti o che venga loro negato un lavoro (perché imporrebbero maggiori costi sanitari al datore di lavoro) a causa dei loro geni.

in un particolare gruppo potrebbe essere proprio la cosa fuori dal controllo di qualcuno.

Nessuno dei sei resoconti di cui sopra costituisce una teoria completa dell'illegittimità della discriminazione. Alcune situazioni che potremmo considerare ovviamente discutibili possono essere colte da alcune teorie, ma non rilevate da altre. Ad esempio, le obiezioni alla discriminazione religiosa non possono basarsi sull'idea che le persone non abbiano controllo sulla propria affiliazione religiosa, ma potrebbero essere supportate facendo riferimento a preoccupazioni relative al pregiudizio o alla mancanza di rispetto.¹⁵¹ O per fare un altro esempio: anche se le azioni dei decisori non sono pregiudizievoli o umilianti, le loro decisioni potrebbero comunque basarsi su caratteristiche irrilevanti: una possibilità che prenderemo in considerazione nella sezione successiva. Sebbene sia improbabile che esista un'unica risposta alla domanda sul perché la discriminazione sia sbagliata, queste teorie sono comunque utili perché evidenziano che spesso dobbiamo considerare molti fattori quando decidiamo se sottoporre gruppi particolari a uno svantaggio relativo sistematico sia moralmente giustificato.

Intenzionalità e discriminazione indiretta

Finora ci siamo concentrati sul perché prendere in considerazione determinate caratteristiche quando si prendono decisioni consequenziali può essere normativamente discutibile. Secondo ciascuna di queste interpretazioni dell'illegittimità della discriminazione, il danno deriva dalla scelta di fare affidamento su queste caratteristiche quando si prendono tali decisioni. Ma cosa succede quando le decisioni non si basano su queste caratteristiche? Eliminare queste caratteristiche dal processo decisionale ne garantisce l'equità?

Un caso semplice è quando il decisore fa intenzionalmente affidamento su proxy per queste caratteristiche (ad esempio, basandosi sul nome di una persona come proxy per il suo genere) al fine di discriminare indirettamente i membri di un gruppo specifico (ad esempio, le donne). Il fatto che tali decisioni non considerino esplicitamente le caratteristiche potrebbe non renderle meno problematiche, dato che il decisore lo fa solo con l'obiettivo di trattare diversamente i membri di questi diversi gruppi. Pertanto, l'illegittimità della discriminazione non si limita al mero utilizzo di determinate caratteristiche nel processo decisionale, ma si estende a qualsiasi tentativo intenzionale di sottoporre i membri di gruppi specifici a un trattamento sistematicamente sfavorevole, anche se ciò viene ottenuto tramite l'uso di deleghe per tali caratteristiche.¹⁵² Tenendo questo in mente, potremmo voler verificare se qualche processo decisionale porta a risultati disparati per i diversi gruppi come un modo per eliminare potenzialmente la discriminazione intenzionale perseguita con l'aiuto di proxy. Se scopriamo disparità nei risultati, potremmo voler verificare se il processo decisionale è stato orchestrato intenzionalmente per raggiungere questo obiettivo, anche se il decisore non sembra prendere esplicitamente in considerazione queste caratteristiche.

Ma che dire delle decisioni che non sono intenzionalmente progettate per discriminare? E se le decisioni non fossero motivate da pregiudizi? Il semplice fatto che un processo decisionale possa portare a risultati abbastanza disparati per gruppi diversi significa che sia ingiusto? Cosa accadrebbe se il decisore potesse offrire qualche motivo per prendere decisioni in questo particolare modo (ad esempio, che il datore di lavoro ha bisogno di persone con credenziali contabili specifiche e che tali credenziali sono detenute più spesso)?

comunemente tra alcuni gruppi rispetto ad altri)?

Possiamo estendere alcuni dei ragionamenti introdotti nella sezione precedente per cercare di rispondere a queste domande. In questo caso, invece di chiederci se i criteri in esame servano semplicemente da proxy per alcune caratteristiche di preoccupazione, potremmo invece chiederci se la scelta dei criteri possa essere giustificata dimostrando che effettivamente servono agli obiettivi dichiarati del decisore.

I processi decisionali che fanno ben poco per raggiungere questi obiettivi, ma che tuttavia sottopongono gruppi specifici a risultati sistematicamente meno favorevoli, sollevano la stessa questione sulla rilevanza che si pone nei casi di discriminazione intenzionale e diretta.

Se i criteri scelti non hanno rilevanza per la decisione in questione, ma risultano in uno svantaggio relativo sistematico per un gruppo specifico, allora fare affidamento su di essi può facilmente diventare funzionalmente equivalente a fare affidamento direttamente sull'appartenenza al gruppo nonostante la sua mancanza di rilevanza per la decisione in questione. In entrambi i casi, il ricorso a criteri irrilevanti è discutibile perché si traduce in uno svantaggio relativo sistematico per gruppi particolari.

Pari opportunità

Che dire di un processo che produce disparità nei risultati ma, di fatto, serve a portare avanti gli obiettivi del decisore? Questo è un caso più difficile su cui ragionare.

Ma prima di farlo, facciamo un passo indietro.

L'uguaglianza di opportunità è un'idea che molti studiosi vedono come l'obiettivo per limitare la discriminazione. L'uguaglianza di opportunità può essere intesa sia in termini ristretti che ampi. La visione ristretta sostiene che dovremmo trattare persone simili in modo simile dato il loro attuale grado di somiglianza. Secondo una visione ampia, dovremmo organizzare la società in modo tale che persone con capacità e ambizioni simili possano ottenere risultati simili. Una posizione a metà strada sostiene che dovremmo trattare persone apparentemente dissimili in modo simile, nella convinzione che il loro attuale grado di dissomiglianza sia il risultato di qualcosa che dovremmo sottovalutare (ad esempio, ingiustizie o sventure passate). Affrontiamo ciascuna visione a turno e vediamo cosa implicherebbe ciascuna riguardo alla domanda di cui sopra.

La visione ristretta

Illustriamo la visione ristretta delle pari opportunità con la nozione di "equità individuale": che le persone che sono simili rispetto a un compito dovrebbero essere trattate in modo simile.¹³⁰ Per ora consideriamo simile come vicinanza media nelle caratteristiche ritenute rilevanti per il compito in questione. mano. L'equità individuale è una nozione comparativa di equità in quanto chiede se esistono differenze nel modo in cui vengono trattate persone simili. Non riguarda direttamente il modo in cui potrebbero essere trattati i membri dei diversi gruppi. Invece, il confronto è tra tutte le persone come individui, non tra i membri di gruppi specifici. Naturalmente, se concordiamo in anticipo che la razza, il genere e così via delle persone sono irrilevanti per il compito da svolgere, allora soddisfare l'equità individuale limiterà anche il grado in cui le persone che differiscono in base a queste caratteristiche riceveranno un trattamento diverso. Ma questo è vero solo per il

misura in cui tali caratteristiche sono ritenute irrilevanti; non è una parte intrinseca della definizione di equità individuale.

L'equità individuale è legata alla coerenza e ad alcune delle preoccupazioni relative all'arbitrarietà che abbiamo esplorato nel capitolo 2. Spesso ci aspettiamo un trattamento coerente in assenza di differenze che sembrerebbero giustificare un trattamento differenziale, soprattutto quando il trattamento determina l'accesso a opportunità importanti. Queste aspettative possono essere così forti che il mancato rispetto di esse provocherà una reazione viscerale: perché non ho ottenuto il trattamento o il risultato desiderato anche se sono apparentemente simile, lungo le dimensioni rilevanti, a qualcuno che lo ha fatto?

Tuttavia, cosa significhi per le persone essere simili non è un dato di fatto. Nella letteratura filosofica, la risposta comune a questa domanda è che le persone dovrebbero essere trattate in modo simile a coloro che sono considerati altrettanto meritori – cioè, che le persone dovrebbero essere giudicate in base alle loro capacità e ambizioni.¹⁵³ Questa comprensione è così È comune che il concetto di pari opportunità – in questa formulazione ristretta – sia spesso considerato sinonimo del concetto di meritocrazia. L'accesso alle risorse e alle opportunità desiderabili dovrebbe essere dettato non dal gruppo sociale a cui qualcuno appartiene, ma piuttosto dalle caratteristiche che sono legittimamente rilevanti per l'istituzione che cerca di portare avanti i propri obiettivi nell'allocazione di tali risorse e opportunità.

Molto dipende da quali decidiamo siano gli obiettivi del processo decisionale. Potrebbe essere definito come la massimizzazione di alcuni risultati di interesse per il decisore, come la prestazione lavorativa. Quando i candidati che si prevede otterranno risultati altrettanto positivi sul lavoro vengono trattati in modo simile nel processo di assunzione, ciò viene spesso interpretato come meritocrazia. Oppure potremmo dire che un'impresa ha un interesse legittimo ad assumere lavoratori con la formazione necessaria per svolgere efficacemente un lavoro, quindi potrebbe assumere solo coloro che hanno completato la formazione necessaria. Se prendere decisioni su questa base porta a tassi di assunzione disomogenei tra i gruppi, nella visione ristretta delle pari opportunità, il decisore è irreprendibile e non ha alcun obbligo di adeguare il processo decisionale.

Ma si noti che i datori di lavoro potrebbero altrettanto facilmente decidere su obiettivi che non hanno alcuna relazione ovvia con ciò che percepiamo come merito, come assumere candidati che probabilmente accetterebbero uno stipendio particolarmente basso. Con questo obiettivo in atto, i decisori avrebbero l'obbligo di trattare allo stesso modo solo le persone che possiedono una sensibilità salariale simile, non coloro che probabilmente otterranno risultati altrettanto positivi sul lavoro. Avrebbe senso descrivere questo come un caso in cui i datori di lavoro garantiscono pari opportunità?¹³⁹ Ciò rivela che la visione ristretta di pari opportunità non detta quale dovrebbe essere il principio normativo costante che determina il modo in cui consideriamo le persone simili; comanda solo che persone simili siano trattate allo stesso modo. Possiede quindi poca sostanza normativa al di là della coerenza. (E anche in questo caso, potrebbero esserci dei limiti pratici a questo principio. Ad esempio, un datore di lavoro potrebbe dover scegliere un solo candidato tra i tanti che si prevede otterranno risultati altrettanto buoni sul lavoro – e quelli che non hanno ottenuto il lavoro potrebbero non obiettare di essere stati trattati ingiustamente.)

Le persone soggette alle decisioni potrebbero avere le proprie idee su come definire la somiglianza, idee che potrebbero essere molto diverse da quelle del decisore.

Ciò potrebbe essere dovuto al fatto che non condividono gli stessi obiettivi del decisore, ma potrebbe anche essere perché credono che ci siano ragioni completamente indipendenti dagli obiettivi del processo decisionale per considerare alcune persone simili. Forse il motivo per cui potremmo considerare due persone diverse come meritevoli di qualche opportunità è perché hanno la stessa probabilità di sfruttare al meglio l'opportunità o perché sono ugualmente bisognosi – non solo ugualmente meritori. In altre parole, potremmo giudicare i candidati come simili perché è probabile che traggano benefici simili dal lavoro, non solo perché è probabile che ottengano risultati altrettanto buoni sul lavoro. In molti casi, la base esatta su cui potremmo considerare le persone come rilevanti simili in un dato contesto può essere piuttosto difficile da articolare per noi perché la nostra concezione di somiglianza potrebbe basarsi su varie considerazioni normative.

La visione ampia

Una visione ampia delle pari opportunità mette da parte le questioni relative all'equità di un dato processo decisionale e si concentra invece sul grado in cui la società nel suo complesso è strutturata per consentire a persone con capacità e ambizioni simili di raggiungere un successo simile. Questa prospettiva è stata sviluppata nel modo più famoso dal filosofo John Rawls sotto lo slogan della “giusta uguaglianza di opportunità”. Per semplificare considerevolmente l'argomento, l'unica ragione difendibile per cui le persone potrebbero sperimentare risultati diversi nel corso della loro vita è se possiedono abilità o ambizioni diverse.¹⁵⁴

Qualunque cosa nella progettazione di particolari istituzioni nella società che impedisca a tali persone di realizzare il proprio potenziale viola questa comprensione più ampia di pari opportunità perché priva le persone ugualmente meritevoli delle stesse possibilità di successo. Ad esempio, una società che non riesce a fornire mezzi a individui con capacità simili nati in circostanze diverse – uno in una famiglia ricca e un altro in una famiglia povera – per raggiungere livelli simili di successo violerebbe questa comprensione delle pari opportunità.

Secondo questo punto di vista, le istituzioni fondamentali che aiutano a coltivare il potenziale delle persone nel corso della loro vita devono essere strutturate in modo da garantire che persone con capacità e ambizioni simili abbiano possibilità simili di ottenere posizioni desiderabili nella società – insieme ai numerosi benefici che ne derivano. tali posizioni. Pertanto, se l'istruzione è un meccanismo importante attraverso il quale il potenziale delle persone può essere promosso, una visione ampia di pari opportunità imporrebbe che le scuole siano finanziate in modo tale che gli studenti con pari capacità e ambizione – provenienti da famiglie ricche o povere – affrontino le stesse prospettive di successo a lungo termine. Pertanto, qualsiasi vantaggio che questi bambini potrebbero ricevere dalle loro famiglie benestanti deve essere compensato da un intervento corrispondente per garantire che questi bambini provenienti da famiglie povere possano prosperare nella stessa misura. Se i bambini più ricchi beneficiano di una base imponibile locale in grado di finanziare una scuola pubblica di alta qualità, allora la società deve mettere in atto politiche che rendano disponibili finanziamenti simili alle scuole pubbliche che educano i bambini più poveri.

Si noti che si tratta di un intervento che mira a eguagliare la qualità dell'istruzione a cui avranno accesso gli studenti ricchi e quelli poveri; non è un intervento

nella politica di ammissione di una particolare scuola. Ciò aiuta a evidenziare il fatto che la visione ampia delle pari opportunità non riguarda realmente l'equità nel processo decisionale; si tratta della progettazione delle istituzioni fondamentali della società, con l'obiettivo di prevenire in primo luogo il verificarsi di ingiuste disuguaglianze. In teoria, il rispetto di tale principio di pari opportunità darebbe come risultato una società senza caste, nella quale nessuno sarebbe permanentemente confinato in una posizione di svantaggio pur avendo il potenziale per avere successo in circostanze diverse.¹⁵⁵ La società sarebbe strutturata per garantire la mobilità sociale per coloro che possiedono la capacità e l'ambizione rilevanti per raggiungere determinati obiettivi.

La visione centrale

Da qualche parte tra i due poli che abbiamo appena esplorato c'è una visione intermedia che è strettamente interessata all'equità del processo decisionale, ma sensibile alle dinamiche attraverso le quali lo svantaggio potrebbe perpetuarsi nella società più in generale. Secondo questo punto di vista, i decisorи hanno l'obbligo di evitare di perpetuare l'ingiustizia.¹⁵⁰ Nello specifico, devono, in una certa misura, trattare persone apparentemente dissimili in modo simile quando le cause di queste diversità sono esse stesse problematiche. Ad esempio, coloro che adottano questo punto di vista potrebbero sostenere che le università non dovrebbero semplicemente classificare i candidati in base alla media dei voti o ai punteggi dei test standardizzati; devono invece valutare i candidati rispetto alle opportunità che sono state loro offerte nel corso della loro infanzia, riconoscendo che il rendimento scolastico e i test standardizzati potrebbero differire in base alle opportunità passate piuttosto che in base alle capacità e ambizioni innate.

Per fare un esempio, lo stato del Texas ha una legge che garantisce l'ammissione alle università finanziate dallo stato a tutti gli studenti che si diplomano nel 10% dei migliori della loro classe di scuola superiore. Questo può essere visto come in linea con la visione centrale. Se l'accesso alle opportunità varia geograficamente, la regola del 10% identifica gli individui con capacità e ambizione senza svantaggiare sistematicamente coloro che hanno avuto la sfortuna di crescere senza accedere a scuole superiori ben finanziate.

Questa visione intermedia differisce da quella ampia nella misura in cui accetta che gli studenti con pari potenziale non riceveranno un'istruzione di pari qualità fino al momento in cui finalmente si iscriveranno all'università. Tuttavia differisce anche dalla visione ristretta nella misura in cui rifiuta di consentire ai college di ignorare ciò che potrebbe spiegare l'attuale dissomiglianza dei candidati nel momento in cui presentano le loro domande. Invece, la visione intermedia suggerisce che vi sia un certo onere per le università nel tentativo di compensare gli svantaggi che alcuni candidati potrebbero aver dovuto affrontare nel corso della loro vita, tanto che potrebbero apparire meno competitivi rispetto ad altri candidati provenienti da contesti più privilegiati.

Di conseguenza, la visione media richiede interventi non a livello di progettazione delle istituzioni, ma a livello di progettazione dei processi decisionali. Suggerisce che per garantire pari opportunità è necessario valutare le persone come sarebbero state se in passato avessero avuto opportunità paragonabili a quelle di altre persone con pari potenziale che cercano l'opportunità attuale. Per certi aspetti, la visione media sembra cercare di realizzare gli obiettivi della visione ampia attraverso a

intervento molto più limitato: mentre la visione ampia sembrerebbe richiedere che i bambini provenienti da famiglie più ricche e quelle più povere abbiano accesso a un'istruzione di pari qualità per tutta la vita, la visione media cerca solo di compensare gli svantaggi sperimentati dagli studenti più poveri rispetto ai loro studenti più ricchi colleghi in specifici momenti decisionali che si ritiene siano particolarmente ad alto rischio – in questo caso, nell'ammissione al college. La visione mediana tende a concentrarsi su questi frangenti perché sembrano essere quelli in cui c'è l'opportunità di alterare notevolmente il corso della vita di un bambino e di consentirgli di realizzare in modo molto più efficace il proprio potenziale.¹⁵⁶ In effetti, questo è spesso il motivo per cui sono percepiti come elevati -posta in gioco.

Mentre gli interventi immaginati dalla visione media potrebbero sembrare più ristretti di quelli della visione ampia perché non richiedono una ristrutturazione più radicale delle istituzioni fondamentali della società, vale la pena notare che gli interventi più discreti della visione media sono progettati per realizzare cambiamento molto più grande di qualsiasi intervento più continuo richiesto da una visione ampia. La visione media prende di mira decisioni specifiche che possono creare un improvviso cambiamento nelle prospettive di vita delle persone, mentre la visione ampia mira a ovviare alla necessità di interventi così drammatici nel processo decisionale garantendo l'uguaglianza durante tutta la vita delle persone.

In altre parole, la visione media richiederà cambiamenti improvvisi e sostanziali in momenti specifici del processo decisionale, mentre la visione ampia richiederà una significativa ridistribuzione delle risorse su base continuativa.

Sebbene la visione mediana proibisca chiaramente di ignorare le ragioni delle differenze di merito tra le persone, non offre una prescrizione chiara su come tenerne conto. Portarlo alla sua logica conclusione comporterebbe interventi che sembrano estremi: potrebbe richiedere di immaginare persone senza gli effetti di secoli di oppressione che loro e i loro antenati potrebbero aver sopportato, suggerendo, ad esempio, che una banca dovrebbe approvare un grosso prestito a qualcuno che in realtà non ha la capacità di ripagarlo. Detto questo, ci sono altre aree del processo decisionale in cui questo punto di vista potrebbe sembrare più ragionevole. Ad esempio, nel mondo del lavoro, potremmo aspettarci che i responsabili delle assunzioni adottino un approccio simile a quello degli addetti alle ammissioni nelle università, valutando le persone in base alle opportunità loro offerte, ignorando alcune differenze nelle qualifiche che potrebbero essere dovute a fattori al di fuori del loro controllo, soprattutto se queste sono qualifiche che il datore di lavoro potrebbe contribuire a coltivare sul posto di lavoro. La visione intermedia ha particolare validità nel caso delle assicurazioni, dove potremmo davvero volere che gli assicuratori ignorino i costi aggiuntivi che probabilmente dovranno affrontare nel fissare il prezzo di una polizza per qualcuno con una costosa condizione preesistente al di fuori del controllo della persona. La misura in cui ci aspettiamo che i decisori si assumano tale responsabilità tende ad essere specifica del contesto e controversa. Torneremo sull'argomento tra breve.

Tensioni tra le diverse visioni

Esiste un evidente conflitto tra l'idea secondo cui i decisori dovrebbero trattare le persone in modo simile a seconda di come appaiono al momento del processo decisionale e l'idea secondo cui i decisori dovrebbero trattare le persone in modo simile a seconda di come appaiono al momento del processo decisionale.

sarebbero apparsi se avessero goduto di privilegi e vantaggi simili a quelli di altri di pari capacità e ambizione.³ Pertanto, una persona che attualmente sembra essere più meritoria rispetto a qualche opportunità potrebbe opporsi se venisse trascurata a favore di qualcuno che a presente sembra meno meritorio, anche se chi prende la decisione ritiene che l'altra persona sarebbe più meritoria della persona che si oppone se entrambi avessero goduto degli stessi privilegi e vantaggi.⁴

Una tensione simile emerge nel modo in cui potremmo cercare di affrontare la discriminazione. La visione ristretta delle pari opportunità suggerisce che il modo per affrontare la discriminazione è garantire che le decisioni siano prese solo sulla base di fattori realmente rilevanti per il compito da svolgere. In altre parole, trattare persone simili in modo simile rispetto al compito dovrebbe, nella maggior parte dei casi, escludere di trattare le persone in modo diverso a seconda del loro genere, razza, ecc., perché è probabile che queste caratteristiche non siano rilevanti per il compito da svolgere. Pertanto, aderire alla visione ristretta delle pari opportunità dovrebbe aiutare a evitare che questi fattori entrino nel processo decisionale. Al contrario, la visione intermedia suggerisce che potremmo voler affrontare la discriminazione considerando esplicitamente queste caratteristiche quando prendiamo decisioni perché è probabile che queste caratteristiche aiuterebbero a spiegare buona parte delle privazioni e degli svantaggi che le persone potrebbero aver dovuto affrontare nel corso del processo delle loro vite. In altre parole, per comprendere come le persone che possiedono queste caratteristiche avrebbero potuto apparire in circostanze controllate, le decisioni devono tenere conto di queste caratteristiche. Anche questo sembra creare un conflitto perché realizzare un impegno verso la visione mediana delle pari opportunità richiede la violazione dei requisiti imposti dalla visione ristretta delle pari opportunità.

John Roemer afferma che queste tensioni si riducono a punti di vista diversi su "quando inizia la competizione" per posizioni desiderabili nella società: a che punto nel corso della nostra vita siamo, in ultima analisi, responsabili di come potremmo confrontarci con gli altri?¹⁵⁷ Dato che abbiamo Senza alcun controllo sulla ricchezza delle famiglie in cui naschiamo o sulla qualità dell'istruzione che potremmo ricevere, potremmo sottovalutare qualsiasi differenza tra le persone che sia dovuta a tali differenze. In altre parole, potremmo dire che non riteniamo ragionevole adottare una visione ristretta

³Queste tensioni sono talvolta espresse come un conflitto tra interventi che riguardano strettamente il processo attraverso il quale vengono prese le decisioni e interventi che riguardano i risultati prodotti da tali decisioni – tra nozioni procedurali e sostanziali di equità.

Evitiamo questa prospettiva e questo linguaggio perché la distinzione si confonde facilmente quando si riconosce che i decisorî a volte potrebbero apportare modifiche al processo decisionale con un occhio al loro effetto sui risultati. Ovviamente, le politiche spesso cercano di evitare determinati risultati distributivi limitando il grado in cui le decisioni possono tenere conto di determinati fattori. Ad esempio, vietando ai datori di lavoro di considerare il sesso o la razza dei candidati, le politiche potrebbero non solo mirare a garantire che tali decisioni siano prese sulla base di informazioni pertinenti, ma mirare a ridurre le disparità nei tassi di assunzione in questo senso.

⁴Si noti che non esiste alcuna tensione evidente tra la visione ristretta e quella ampia perché è ampia Questa visione richiederebbe che persone con pari capacità e ambizione abbiano ricevuto opportunità simili in una fase molto precedente della loro vita, apparentemente già simili nel momento in cui arrivano al momento attuale del processo decisionale. Naturalmente, questo sarebbe vero solo in una società impegnata in una significativa quantità di redistribuzione – una prospettiva alla quale i sostenitori della visione ristretta delle pari opportunità potrebbero fortemente opporsi.

di pari opportunità nella valutazione dei candidati al college perché molte delle differenze rilevanti tra i candidati potrebbero non essere emerse da una competizione leale. Al contrario, potremmo pensare che i datori di lavoro siano giustificati nel valutare i candidati per un posto di lavoro, soprattutto quelli per ruoli più senior che richiedono molti anni di esperienza, in base alle capacità e all'ambizione che hanno dimostrato nel corso della loro carriera. Potremmo cioè essere d'accordo sul fatto che una visione ristretta delle pari opportunità sia appropriata in questo caso perché le persone che hanno fatto carriera bene hanno avuto un'equa possibilità di coltivare le proprie capacità e dimostrare la propria ambizione. Le tensioni sorgono quando c'è disaccordo su dove avviene questa transizione nella vita delle persone. Qualcuno che è stato scavalcato a favore di un'altra persona che sembra meno meritoria potrebbe considerarlo ingiusto perché pensa che qualunque differenza esista tra loro due sia emersa durante un periodo di leale competizione tra loro. Il decisore potrebbe non essere d'accordo, ritenendo che le differenze in realtà siano dovute ai vantaggi che l'individuo trascurato ha maturato durante un periodo precedente all'inizio di una competizione leale.

Tabella 4.1: Alcune differenze chiave tra i tre punti di vista sulle pari opportunità.

	Obiettivo	Intervento punto	Chi sostiene il costo del miglioramento dei gruppi storicamente svantaggiati
Vista ristretta	Garantire che le persone che hanno le stesse qualifiche per un'opportunità abbiano le stesse possibilità di ottenerla	Processo decisionale Nessuno ⁵	
Vista centrale	Differenze di sconto dovute a ingiustizie passate che spiegano le attuali differenze nelle qualifiche	Processo decisionale, in particolare opportunità di vita critiche decisionali)	Decision maker (che può trasferire i costi ai soggetti
Visione ampia	Garantire che persone con pari capacità e ambizione siano in grado di realizzare il proprio potenziale altrettanto bene	Governo, su base continuativa	Contribuenti

⁵ Gli interventi secondo la visione ristretta includono l'adozione di criteri decisionali più rilevanti e la raccolta di più dati sugli argomenti decisionali che aiutano a prendere decisioni più accurate. Questi interventi, in definitiva, avvantaggiano gli obiettivi del decisore (nella misura in cui tali obiettivi sono moralmente legittimi), quindi non li consideriamo come costi per il decisore.

Merito e deserto

Anche se accettiamo una visione media e ampia delle pari opportunità, potremmo aver bisogno di alcuni principi normativi che ci permettano di decidere fino a che punto i decisori e il governo devono spingersi nel tentativo di contrastare la disuguaglianza. Più concretamente, quali differenze tra le persone giustificano effettivamente le differenze nei risultati, se ce ne sono? Finora abbiamo avuto la tendenza a rispondere a questa domanda descrivendo quelle differenze che non possono o non devono servire da giustificazione per le differenze nei risultati. Ma vale anche la pena riflettere più a fondo sui principi che sembrano consentire – o forse addirittura richiedere – queste differenze nei risultati. In questa sezione discuteremo due principi che aiutano a rispondere a questa domanda. Il primo, che ha già avuto un ruolo importante nella nostra discussione, è il principio del merito. Il secondo è il principio del deserto, ovvero ciò che è meritato.

Il merito gioca un ruolo importante in tutte e tre le visioni delle pari opportunità. Nella nostra discussione sulla visione ristretta delle pari opportunità, il merito è un modo per stabilire in che modo le persone sono simili e quindi chi dovrebbe essere trattato in modo simile. In una visione ampia, il merito – sotto forma di abilità e ambizioni – consente la possibilità che persone con meriti diversi non raggiungano risultati comparabili nella vita, ma determina anche la quantità di sostegno che deve essere fornita a persone con gli stessi meriti. potenziale come i loro coetanei più privilegiati, ma che non sarebbero in grado di realizzare il proprio potenziale in modo altrettanto efficace in assenza di tale sostegno. Il merito è fondamentale per l'idea che esista un obbligo morale di aiutare le persone a realizzare il proprio potenziale, ma nessun obbligo di andare oltre. Infine, il merito gioca un ruolo simile nella visione centrale in quanto ci si aspetta che i decisori valutino le persone in base a quanto sarebbero state meritorie in circostanze controllate. Intesi in questo modo, tutti e tre i punti di vista sono forse più simili di quanto potrebbero sembrare a prima vista: ciascuno chiede che persone con meriti simili abbiano le stesse possibilità di successo.

Ma cos'è esattamente il merito? Il merito non è una proprietà oggettiva posseduta da un dato individuo. Invece, il merito riguarda le qualità possedute da una persona specifica che ci si aspetta contribuiscano a raggiungere l'obiettivo dell'istituzione che offre l'opportunità ricercata.¹⁵³ Pertanto, ciò che rende meritorio un particolare candidato a un posto di lavoro è la probabilità che quel candidato faccia progressi. gli obiettivi del datore di lavoro. Sebbene sia forte la tentazione di pensare che esistano risposte universali a ciò che rende un candidato più o meno meritorio di altri (ad esempio, quanto è intelligente , quanto lavora duramente, ecc.), non è così. Invece, il merito, da questo punto di vista, è puramente una funzione di ciò che un datore di lavoro considera rilevante per raggiungere i propri obiettivi, qualunque essi siano. E diversi datori di lavoro potrebbero avere obiettivi molto diversi e idee molto diverse su cosa potrebbe fare di più per aiutarli a realizzarli. Gli obiettivi del datore di lavoro potrebbero non essere gli obiettivi che il candidato al lavoro vorrebbe che fossero o quelli che gli esterni vorrebbero che fossero. La concezione del merito di altri potrebbe differire da quella dei datori di lavoro perché semplicemente hanno idee diverse sugli obiettivi che i datori di lavoro dovrebbero avere in primo luogo.

La soggettività di questa visione del merito sembra essere in conflitto con le precedenti discussioni su capacità e ambizioni, che vengono presentate come proprietà universali

che non sono legati agli obiettivi di una particolare istituzione che fornisce l' opportunità ricercata. Nel suggerire che persone con capacità e ambizioni simili dovrebbero avere possibilità simili di ottenere posizioni desiderabili nella società, sembra esserci la convinzione implicita che le persone possano essere paragonate in base ai loro meriti, indipendentemente dall'opportunità in questione. Ciò riflette il fatto che spesso esistono convinzioni ampiamente condivise e ben radicate sulla rilevanza di determinati criteri nel determinare se qualcuno merita una particolare opportunità. Si ritiene comunemente che un datore di lavoro che valuta le persone in base alla loro intelligenza e operosità valuti le persone in base al loro merito perché queste sono le proprietà che si possono tranquillamente assumere per aiutare il datore di lavoro a promuovere i propri interessi.

Ma non vi è alcun motivo per cui queste debbano essere le caratteristiche in base alle quali i candidati devono essere valutati per garantire che le decisioni del datore di lavoro siano basate sul merito.

Questa osservazione anticipa una nozione correlata: deserto. A differenza del merito, il merito non è legato a quanto bene una persona potrebbe contribuire a promuovere gli interessi del decisore , ma a quanto bene una persona si è comportata rispetto alle dimensioni in cui ci si aspetta che un decisore valuti le persone. Ad esempio, potremmo dire che una persona che ha lavorato diligentemente durante tutta la scuola per ottenere voti alti è più meritaria rispetto a un'opportunità di lavoro rispetto a una persona che ha saltato le lezioni e ha ricevuto voti medi, anche se entrambe le persone hanno probabilità di raggiungere gli obiettivi. di un datore di lavoro altrettanto bene. In questo contesto, le persone meritano determinate opportunità poiché potrebbero avere buone ragioni per credere che determinati investimenti li aiuterebbero ad accedere all'opportunità ricercata. In altre parole, i decisori hanno l'obbligo di offrire opportunità alle persone che hanno intrapreso azioni per le quali meritano di essere ricompensate.

Questo principio può aiutare a spiegare perché crediamo che le persone che intendono creare una famiglia dovrebbero avere le stesse possibilità di assicurarsi un lavoro di coloro che non lo fanno quando dimostrano pari capacità e ambizione, anche se per avviare una famiglia i dipendenti devono andare in congedo per periodi prolungati e anche se ciò aumenta la probabilità che i dipendenti lascino il posto di lavoro, imponendo così ai datori di lavoro costi che potrebbero altrimenti essere evitati. Mentre l'obiettivo del datore di lavoro potrebbe essere quello di assumere persone che probabilmente lavoreranno diligentemente senza interruzioni e che probabilmente resteranno in azienda a tempo indeterminato, selezionare tra i candidati su questa base potrebbe indurre il datore di lavoro a sfavorire i candidati che meritano di essere assunti alla luce delle loro qualità. capacità e ambizione. In particolare, le aspiranti madri, che sono più propense delle loro coetanee a prendersi una pausa o a lasciare il lavoro per fondare una famiglia (a causa delle norme di genere radicate sulla divisione delle responsabilità genitoriali), non dovrebbero essere trascurate a favore di altre se hanno tutti fatto gli stessi investimenti per prepararsi a candidarsi per queste posizioni. Il principio del deserto afferma che coloro che hanno maturato le competenze necessarie per avere successo sul lavoro dovrebbero essere tutti valutati allo stesso modo, indipendentemente dalle differenze nella probabilità che i candidati debbano prendersi una pausa o rinunciare al proprio lavoro.

Le discussioni sul merito e sul merito aiutano anche a evidenziare che possono esserci giustificazioni molto diverse per i vincoli o le richieste che potremmo imporre ai decisori. In alcuni casi, potremmo sostenere che le persone sono semplicemente moralmente

diritto ad un determinato trattamento. Ad esempio, potremmo dire che è sbagliato ritenere le persone responsabili di caratteristiche personali sulle quali non hanno alcun controllo, anche se farlo sarebbe nell'interesse razionale di un decisore.

Allo stesso modo, potremmo dire che alle persone sono dovute determinate opportunità alla luce delle loro capacità e ambizioni, anche se il decisore preferirebbe giudicare le persone su basi diverse. Questi sono ciò che i filosofi chiamano argomenti deontologici: ragioni morali per cui alcune azioni sono preferibili ad altre indipendentemente dalle conseguenze di queste azioni. Dobbiamo sottovalutare gli effetti della sfortuna e tenere conto del merito perché questo è ciò che richiede l'equità.

Al contrario, potremmo sostenere che il modo in cui i decisori trattano le persone dovrebbe essere dettato dalle conseguenze di tali azioni. Ad esempio, potremmo dire che il processo decisionale basato sul merito è giustificato sulla base del fatto che l'assegnazione delle opportunità in base al merito aiuta a promuovere l'interesse della società, non solo dell'individuo che cerca una particolare opportunità o del decisore che offre l'opportunità. Assumere sulla base delle capacità e dell'ambizione può avere la conseguenza di migliorare il benessere generale se ciò significa che le persone particolarmente ben preparate a intraprendere determinate attività hanno maggiori probabilità di ottenere l'opportunità di farlo. Il processo decisionale basato sul merito è quindi giustificato perché mette a frutto i talenti degli individui a vantaggio collettivo della società, non perché ogni dato individuo abbia moralmente diritto a una particolare opportunità alla luce dei suoi meriti. Potremmo inoltre sostenere che il trattamento differenziato sulla base del merito incentiva e premia gli investimenti produttivi a vantaggio di tutta la società.

Naturalmente, ci sono anche argomenti consequenzialisti a favore di interventi volti a sollevare coloro che hanno subito svantaggi o discriminazioni in passato. Ad esempio, la società soffre nel complesso quando ai membri di gruppi specifici viene negata l'opportunità di realizzare il loro vero potenziale perché rinunciano ai benefici collettivi che potrebbero essere apportati dai contributi di tali gruppi.

Il costo dell'equità

Diversi punti di vista sull'uguaglianza di opportunità – così come le nozioni di merito e merito da cui dipendono tali punti di vista – assegnano la responsabilità e i costi associati nella gestione dell'ingiustizia in modo molto diverso. In particolare, la visione mediana attribuisce l'onere ai singoli decisori e alle istituzioni specifiche, indipendentemente dalla velocità con cui o dalla misura in cui una persona è in grado di realizzare il proprio potenziale. Ad esempio, potremmo aspettarci che le università sostengano alcuni costi iniziali nell'ammettere studenti provenienti da contesti meno privilegiati perché le università potrebbero dover investire risorse aggiuntive per aiutare questi studenti a stare al passo con i loro coetanei più privilegiati. Ciò potrebbe assumere la forma di fornire lezioni durante l'estate che precede l'inizio dei programmi universitari formali. Oppure potrebbe assumere la forma di progettare corsi introduttivi senza dare per scontate molte conoscenze di base, che potrebbero dedicare un po' di tempo a rivedere materiale familiare agli studenti delle scuole superiori con maggiori finanziamenti, ma forse nuovo per coloro che provengono da scuole con minori finanziamenti. Le università potrebbero

investono anche in programmi che cercano di limitare il grado in cui le disuguaglianze esistenti tra gli studenti prima di iscriversi all'università si trasmettono durante le loro esperienze universitarie. Ad esempio, le università potrebbero offrire borse di studio finanziarie agli studenti più poveri con l'obiettivo di consentire loro di evitare di dover lavorare per mantenersi, consentendo così a questi studenti di dedicare allo studio una quantità di tempo simile a quella dei loro coetanei più privilegiati. Tali borse di studio potrebbero anche aiutare a evitare di gravare gli studenti più poveri con debiti significativi, che potrebbero sopprimere i guadagni futuri e influenzare negativamente le scelte di carriera – oneri che gli studenti più ricchi senza debiti significativi non devono affrontare. Gli interventi in questo senso offuscano la distinzione tra la visione media e ampia dell'uguaglianza di opportunità perché sembrano mirati non a rimediare ad alcune ingiuste disuguaglianze del passato, ma a prevenire il riemergere di un'ingiusta disuguaglianza. Il capitolo 8 tratterà tali interventi in modo più approfondito.

Nonostante questi sforzi, le università potrebbero scoprire che i loro investimenti in questo tipo di interventi potrebbero richiedere molti anni per essere ripagati: gli studenti provenienti da contesti meno privilegiati potrebbero seguire i loro coetanei provenienti da contesti più privilegiati nei voti che ottengono nel corso dei loro studi universitari. carriere, ma alla fine raggiungono un successo comparabile una volta entrati nel mercato del lavoro.

Allo stesso modo, i datori di lavoro che assumono candidati di cui riconoscono un grande potenziale, ma che necessitano anche di ulteriore supporto, potrebbero non essere i datori di lavoro che beneficiano di tali investimenti. I dipendenti potrebbero accettare un altro lavoro prima che il datore di lavoro originario ritenga di aver recuperato il proprio investimento. Questo è un aspetto importante della visione mediana delle pari opportunità perché evidenzia che potrebbe non essere sempre nell'interesse razionale dei decisori comportarsi in questi modi. (Tuttavia, questo potrebbe anche tagliare nel senso opposto: un decisore non vincolato potrebbe scartare qualcuno che sembra meritorio perché riconosce che la persona ha beneficiato della buona fortuna - e quindi è priva dell'abilità o dell'ambizione che sta effettivamente cercando. per.)

La visione mediana non è quindi semplicemente una tesi secondo cui i decisori devono badare ai propri interessi personali a lungo termine; si sostiene che alcune istituzioni siano gli attori giusti per sostenere dei costi al servizio della correzione della disuguaglianza e dell'ingiustizia, anche se non vi è alcuna garanzia di ottenere una ricompensa di valore almeno equivalente .

Ciò contrasta con la visione ampia delle pari opportunità, secondo cui il governo è considerato l'attore appropriato per facilitare la redistribuzione necessaria a compensare le ingiuste disparità, probabilmente attraverso la tassazione diretta e i trasferimenti. Secondo la visione ampia, il governo – vale a dire tutti coloro che pagano le tasse al governo – ha l'onere di contrastare i vantaggi di cui altrimenti godrebbero, ad esempio, gli studenti provenienti da contesti più privilegiati. Nella misura in cui sono necessari interventi da parte dei datori di lavoro o di altre istituzioni, il governo dovrebbe sovvenzionare i loro sforzi con i soldi delle tasse. Al contrario, la visione media attribuisce questo onere a specifici decisori per compensare gli svantaggi che le persone hanno già sperimentato.

Tutto ciò suggerisce una serie di domande difficili: in che misura il peso della discriminazione passata dovrebbe ricadere sui singoli decisori? In che orari

dovremmo tentare di correggere gli effetti dell'ingiustizia storica? Ed è possibile compensare il risultato cumulativo delle migliaia di momenti in cui le persone si trattano in modo diverso nel corso della vita? Torneremo su queste domande nel capitolo 6 quando considereremo, da un punto di vista legale e pratico, chi potremmo considerare nella posizione migliore per sostenere questi costi.

Collegamento di nozioni statistiche e morali di equità

Cerchiamo ora di collegare alcune delle nozioni morali che abbiamo discusso finora in questo capitolo ai criteri statistici del Capitolo 3. Naturalmente, molti dei concetti di questo capitolo, come se un soggetto decisionale abbia il controllo su un attributo utilizzato per il processo decisionale, non può essere espresso nel linguaggio statistico delle probabilità e delle aspettative condizionate. Inoltre, anche per nozioni che sembrano tradursi in condizioni statistiche, ribadiamo la nostra consueta nota di cautela secondo cui i criteri statistici da soli non possono certificare che un sistema sia giusto. Tanto per fare un esempio, i criteri del capitolo 3 non variano in base ai tassi di domanda dei diversi gruppi. Ad esempio, se il 50% dei richiedenti un prestito di un particolare gruppo decidesse di non fare domanda per qualche motivo, un classificatore che soddisfaceva l'indipendenza/parità demografica prima della modifica dei tassi di richiesta continuerebbe a soddisfare l'indipendenza/parità demografica dopo la modifica. Lo stesso vale per la parità di sufficienza/calibrazione e di separazione/tasso di errore. Tuttavia, non considereremmo giusto una banca, un datore di lavoro o un'altra istituzione se scoraggiasse le richieste di determinate persone o gruppi. Ciò è legato alla trappola del framing di Selbst et al.: un "fallimento nel modellare l'intero

sistema su cui verrà applicata [l'equità]".¹⁵⁸ Ma dobbiamo anche resistere all'estremo opposto, ovvero l'idea che i criteri statistici non abbiano contenuto normativo. Riteniamo che i criteri statistici siano un aspetto di ciò che significa per un sistema sociotecnico essere giusto e, combinati con le tutele procedurali, possono aiutarci a raggiungere diversi obiettivi morali.

Parità demografica

Tolte queste avvertenze, cominciamo con un criterio statistico relativamente semplice: la parità demografica. Ha un rapporto tenue ma evidente con la visione ampia delle pari opportunità nella misura in cui mira a uniformare i risultati. La somiglianza di alto livello tra i due è l'idea di distribuzione proporzionale delle risorse. Ma le nozioni morali non corrispondono mai esattamente ai criteri tecnici. Consideriamo le differenze tra loro come un modo per comprendere la relazione.

La visione ampia delle pari opportunità riguarda i risultati della vita delle persone (come la ricchezza) piuttosto che i momenti discreti del processo decisionale. Tuttavia, possiamo sperare che imporre una qualche nozione di uguaglianza nelle decisioni che influenzano l'esito degli interessi (come i posti di lavoro nel caso della ricchezza) porti all'uguaglianza nel risultato corrispondente. Empiricamente, tuttavia, non è affatto chiaro che imporre l'uguaglianza nel breve termine porterà all'uguaglianza nel lungo termine. In effetti, il lavoro teorico ha dimostrato che non è sempre così.¹⁵⁹

Inoltre, l'uguaglianza dei risultati – cioè l'applicazione di eguali risultati nella vita – ignora le differenze di abilità e ambizione tra le persone che potrebbero ragionevolmente giustificare le differenze nei risultati. Questa è anche la critica più comune all'uguaglianza dei risultati: esclude la particolare comprensione del processo decisionale basato sul merito che è alla base dell'applicazione dell'apprendimento automatico in molti contesti.

Anche se questa è spesso considerata un'obiezione fatale all'uguaglianza dei risultati, la critica perde gran parte della sua forza se applicata alla parità demografica. Per cercare di giustificare la parità demografica nonostante le differenze individuali in termini di capacità e ambizione, riconosciamo tali differenze ma sosteniamo che queste si annullano a livello di gruppi; quindi, mentre le decisioni prese sugli individui possono essere in sintonia con le differenze tra loro, richiediamo che i benefici e gli oneri di tali decisioni siano equamente distribuiti tra i gruppi, in media.

Ma a quali gruppi dovremmo prestare attenzione? Come in precedenza, prestiamo particolare attenzione alle differenze di gruppo quando riteniamo che siano particolarmente probabili a sorgere a causa di condizioni storiche ingiuste o ad aggravarsi nel tempo. Questi corrispondono agli assi lungo i quali la società si è storicamente ed è attualmente stratificata. In questa prospettiva, potremmo preoccuparci dell'uguaglianza di risultati non solo fine a se stessa, ma anche perché la disuguaglianza di risultati è un buon indicatore del fatto che potrebbe esserci disuguaglianza di opportunità nel senso ampio del termine.¹⁶⁰ In altre parole, alcune disuguaglianze i risultati potrebbero non essersi verificati se non ci fossero state in passato alcune disuguaglianze di opportunità.

Esistono molti altri divari tra la parità demografica e l'uguaglianza dei risultati. Ne citiamo solo un altro: non tutti i soggetti decisionali (e i gruppi) possono valorizzare la risorsa allo stesso modo. Le pubblicità mirate possono essere utili per gli individui più ricchi informandoli su ciò che i loro soldi possono comprare, ma sfruttano le insicurezze economiche degli individui più poveri (ad esempio i prestiti con anticipo sullo stipendio¹⁶¹). La polizia può essere utile per alcune comunità ma gravare su altre, a seconda dei pregiudizi degli agenti di polizia. In questi casi, i risultati effettivi – benefici e danni – possono essere molto diversi nonostante la parità statistica nell'allocazione.

Calibrazione

Ricordiamo dal capitolo 3 che se l'appartenenza al gruppo è codificata in modo ridondante nelle caratteristiche, il che è più o meno vero in set di dati sufficientemente ricchi, allora la calibrazione è una conseguenza dell'apprendimento supervisionato non vincolato. Pertanto, può essere raggiunto senza prestare esplicita attenzione all'appartenenza al gruppo. In altre parole, imporre la calibrazione come requisito non è un grande intervento.

Tuttavia, il concetto ha un fascino intuitivo: se un punteggio è calibrato per gruppo, allora sappiamo che un valore di punteggio (ad esempio, il 10% di rischio di default) indica lo stesso tasso di risultati positivi (ad esempio, tasso di default) in tutti i gruppi. Allo stesso modo, ha una certa utilità diagnostica dal punto di vista dell'equità, come ad esempio segnalare la discriminazione "irrazionale". Se il classificatore codifica esplicitamente una preferenza per un gruppo o un pregiudizio nei confronti di un altro (o un decisore umano esercita tale preferenza o pregiudizio), la distribuzione risultante non sarà calibrata per gruppo.

La calibrazione può anche essere vista come un controllo di integrità per l'ottimizzazione. Proprio perché la calibrazione è implicita nell'ottimizzazione non vincolata, possiamo rilevare i fallimenti di ottimizzazione derivanti da violazioni della calibrazione. Ma questo è tutto: un controllo di integrità. Un modello può essere estremamente impreciso e tuttavia soddisfare la calibrazione. In effetti, un modello senza potere discriminativo che restituisce sempre semplicemente il risultato medio della popolazione è perfettamente calibrato. Un modello che è estremamente accurato per un gruppo (ottimale come definito nel Capitolo 3) e che prevede sempre la media per un altro gruppo è anche perfettamente calibrato.

La calibrazione per gruppo si adatta a una visione ristretta delle pari opportunità. Supponiamo che un decisore utilizzi solo le caratteristiche ritenute rilevanti, mentre l'appartenenza al gruppo è ritenuta irrilevante. Quindi la calibrazione per gruppo dice che il decisore non considera l'appartenenza al gruppo oltre la misura in cui è codificata in caratteristiche rilevanti per il compito. La decisione può giustificare le differenze di gruppo nei risultati facendo appello alle differenze nelle caratteristiche rilevanti.

È necessaria una giustificazione normativa non banale per violare la calibrazione nei modelli utilizzati per il processo decisionale. Abbiamo discusso molte di queste giustificazioni, come la convinzione che il rischio derivi in parte da fattori di cui il soggetto decisionale non dovrebbe essere ritenuto responsabile.

Il criterio di somiglianza

Torniamo al criterio di somiglianza: trattare persone simili in modo simile. Come abbiamo discusso, la sostanza normativa di questa nozione si riduce in gran parte a ciò che intendiamo per simile. Una visione comune è pensarla come una vicinanza rispetto alle caratteristiche che si riferiscono alle qualifiche per il compito da svolgere, interpretando le caratteristiche per valore nominale.

Per tradurre questo in una nozione tecnica, possiamo immaginare di definire una funzione o metrica di somiglianza specifica per un compito tra due vettori di caratteristiche che rappresentano individui. Possiamo quindi insistere sul fatto che per due individui sufficientemente simili, le decisioni che ricevono saranno corrispondentemente simili. Lo chiamiamo criterio di somiglianza. Questa nozione è stata precisata e analizzata da Cynthia Dwork, Moritz Hardt, Toniann Pitassi et al.¹³⁰ Una volta che abbiamo una metrica, possiamo risolvere un problema di ottimizzazione vincolata. L'obiettivo di ottimizzazione è quello usuale (ad esempio, minimizzare la differenza tra la prestazione lavorativa prevista e quella osservata) e il criterio di somiglianza è il vincolo.

Possiamo illustrare questo approccio nel contesto della pubblicità comportamentale online . La nostra discussione presuppone che consideriamo gli annunci pubblicitari come un'assegnazione di accesso a opportunità (ad esempio, mirando a opportunità di lavoro o offerte di credito). Le reti pubblicitarie raccolgono informazioni demografiche sugli individui, come la cronologia di navigazione, la posizione geografica e il comportamento di acquisto, e le utilizzano per assegnare una persona a una delle poche dozzine di segmenti. I segmenti hanno nomi come "White Picket Fences", una categoria di mercato con un reddito familiare medio di poco superiore a \$ 50.000, di età compresa tra 25 e 44 anni con figli, con una certa istruzione universitaria, ecc. Gli individui in un segmento sono considerati simili per scopi di marketing e gli inserzionisti è consentito indirizzare gli annunci solo a livello di segmenti e non di individui.

Ciò riflette la visione ristretta delle pari opportunità. Se due individui differiscono solo su dimensioni ritenute irrilevanti per gli interessi commerciali dell'inserzionista, ad esempio la religione, si troveranno nello stesso segmento e quindi ci si aspetta che vedano gli stessi annunci. D'altra parte, se alcune persone o gruppi sociali hanno goduto di vantaggi nel corso della loro vita che hanno consentito loro di raggiungere un certo livello di reddito, allora il criterio di somiglianza consente ai benefici di tali vantaggi di riflettersi negli annunci che vedono.

La pubblicità mirata è un dominio particolarmente buono per applicare queste idee. Esiste un intermediario, la rete pubblicitaria, che comprime i vettori delle funzionalità in categorie (ad esempio, segmenti di annunci) ed espone solo queste categorie agli inserzionisti, anziché consentire direttamente agli inserzionisti di rivolgersi agli individui. La rete pubblicitaria dovrebbe costruire i segmenti in modo tale che i membri simili per scopi pubblicitari debbano trovarsi nello stesso segmento. Ad esempio, non sarebbe accettabile includere un segmento corrispondente alla disabilità, perché la disabilità non è un criterio di targeting rilevante per la stragrande maggioranza dei tipi di annunci. In ambiti diversi dalla pubblicità mirata, ad esempio le ammissioni all'università, applicare queste idee è più impegnativo. In assenza di un intermediario come la rete pubblicitaria, spetta a ciascun decisore fornire trasparenza nella propria metrica di somiglianza.

Questa interpretazione restrittiva del criterio di somiglianza si collega ad altre definizioni formali di equità individuale, come la nozione di equità meritocratica nel contesto dell'apprendimento bandito.¹⁶² Il contenuto normativo del criterio di somiglianza, tuttavia, si estende oltre la visione ristretta dell'uguaglianza dei opportunità se ampliamo i principi da cui costruiamo una metrica di somiglianza. Ad esempio, la nozione di somiglianza potrebbe modificare esplicitamente le caratteristiche in base alla considerazione delle ingiustizie e degli svantaggi passati. Potremmo essere d'accordo fin dall'inizio sul fatto che un punteggio del test SAT di 1200 in determinate circostanze corrisponde a un punteggio di 1400 in condizioni di fondo più favorevoli.

Confronti come questi sono strettamente legati alla definizione formale di uguaglianza di opportunità di Roemer.¹⁵⁷ Roemer immagina una partizione della popolazione in tipologie basate su caratteristiche "facilmente osservabili e non manipolabili" che si riferiscono a "circostanze differenti di individui per le quali crediamo che non dovrebbero essere ritenuto responsabile". La definizione formale confronta quindi gli individui che spendono lo stesso quantile di impegno rispetto alla loro tipologia.

Randomizzazione, soglia ed equità

Se pensiamo di applicare il criterio di somiglianza a un compito come l'assunzione, ci imbattiamo in un altro problema: coppie di candidati estremamente simili possono trovarsi su lati opposti della soglia di punteggio, perché dobbiamo tracciare una linea da qualche parte. Ciò violerebbe il criterio di somiglianza. Un modo per superare questo problema è insistere affinché il classificatore sia randomizzato.

La randomizzazione a volte offende intuizioni morali profondamente radicate, soprattutto in ambiti come la giustizia penale, evocando lo spettro di decisioni prese sulla base del lancio di una moneta. Ma ci sono diverse ragioni per cui in alcuni casi la randomizzazione potrebbe non solo essere accettabile ma necessaria per ragioni di equità (oltre al fatto che

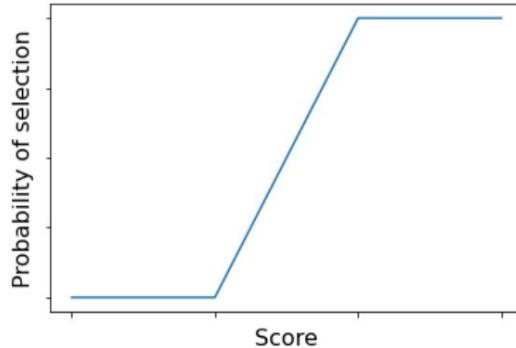


Figura 4.1: Un classificatore randomizzato. Solo i classificatori randomizzati possono soddisfare il criterio di somiglianza. Due individui simili avrebbero punteggi simili e quindi probabilità di selezione simili.

ci permette di trattare persone simili in modo simile, almeno in senso probabilistico).

Infatti, Ronen Perry e Tal Zarsky presentano numerosi esempi di casi in cui la legge richiede che le decisioni consequenziali siano basate su lotterie.¹⁶³

Per comprendere la giustificazione del processo decisionale randomizzato, dobbiamo riconoscere che la casualità mirata e controllata con precisione non è la stessa cosa dell'arbitrarietà o capricciosità. Supponiamo che valgano queste tre condizioni: una risorsa da allocare è indivisibile, ci sono meno unità di essa rispetto ai richiedenti e non c'è nulla che dia diritto alla risorsa a un richiedente più o meno degli altri richiedenti.

Allora la randomizzazione potrebbe essere l'unico modo egualitario per rompere il pareggio. Questo è il principio alla base delle lotterie per l'assegnazione di alloggi pubblici a basso costo e di visti di immigrazione. Lo stesso principio si applica agli oneri piuttosto che alle risorse valutate, come si vede nella selezione casuale di persone per compiti di

giuria o controlli fiscali.¹⁶³ Ma il punto centrale dell'utilizzo dell'apprendimento automatico è che esiste un modo per classificare le richieste dei ricorrenti, quindi gli scenari che ci interessano sono più complicati degli esempi precedenti. La complicazione è che esiste un conflitto tra gli obiettivi di trattare persone simili in modo simile (che richiede la randomizzazione) e di ridurre al minimo l'imprevedibilità nella decisione (che richiede di evitare la randomizzazione).

Una distinzione fondamentale che influisce sulla legittimità della randomizzazione è se esistono opportunità equivalenti per le quali un richiedente potrebbe essere idoneo. Ciò è generalmente vero nel caso di assunzione o prestito, ma non nel caso di custodia cautelare. La randomizzazione è più giustificabile nel primo caso perché evita il problema che un richiedente rimanga perennemente appena al di sotto della soglia di selezione. Se viene utilizzata la randomizzazione, un candidato ragionevolmente qualificato ma non eccezionale potrebbe dover candidarsi per diversi lavori, ma alla fine ne otterrà uno.¹⁶⁴

Un altro modo per evitare il problema che candidati simili cadano su lati opposti di un limite è riprogettare il sistema in modo che le decisioni non siano binarie. Anche questo è più facile per alcune istituzioni rispetto ad altre. Un creditore può tenere conto dei diversi livelli di rischio adattando il tasso di interesse per un prestito anziché rifiutare il prestito

del tutto. Al contrario, nel sistema di giustizia penale è incorporata una nozione binaria di determinazione della colpa.⁶ Questa situazione non è facile da cambiare. Si noti che le determinazioni della colpevolezza non sono previsioni; hanno lo scopo di riflettere una verità binaria e l'obiettivo del sistema di giustizia penale è scoprirla.

I fondamenti normativi della parità del tasso di errore

Delle tre principali famiglie di criteri statistici nel Capitolo 3, abbiamo discusso indipendenza / parità demografica e sufficienza/calibrazione, lasciando la separazione /parità del tasso di errore. La parità del tasso di errore è il criterio più difficile da collegare rigorosamente a qualsiasi nozione morale. Allo stesso tempo, è innegabile che attinga a un'intuizione di equità ampiamente condivisa. Lo studio di ProPublica sul sistema di previsione del rischio penale COMPAS è stato così potente perché ha scoperto che gli imputati neri avevano "il doppio del tasso di falsi positivi" degli imputati bianchi.¹³³ Ma

non esiste una giustificazione diretta per questa intuizione, che ha portato a raggiungere la parità del tasso di errore un argomento di acceso dibattito.^{165, 166} Basandosi su questo studio, forniamo la nostra visione del motivo per cui dovremmo preoccuparci della parità del tasso di errore.

Assumeremo un problema di allocazione delle risorse basato sulla previsione, come il prestito, che presenta un sostanziale grado di incertezza intrinseca rispetto alla prevedibilità dei risultati. Al contrario, la disparità nel tasso di errore spesso emerge in problemi di percezione come il riconoscimento facciale o il rilevamento del linguaggio dove c'è poca o nessuna incertezza intrinseca.^{167, 168} La differenza normativa cruciale è che nel riconoscimento facciale, nel rilevamento del linguaggio e applicazioni simili, c'è non vi è alcuna nozione di differenza di qualificazione tra individui che possa potenzialmente giustificare un trattamento dissimile.

Pertanto, partendo dal presupposto che un'errata classificazione impone un costo in materia, è molto più semplice giustificare il motivo per cui tassi di errore disuguali sono problematici.

Un'altra osservazione per preparare il terreno: il significato morale del tasso di errore è asimmetrico. Un tipo di errore corrisponde, grosso modo, a un'ingiusta negazione (della libertà o di un'opportunità) e l'altro corrisponde a un trattamento eccessivamente indulgente. Nella maggior parte dei settori, il primo tipo è molto più significativo dal punto di vista normativo rispetto al secondo. Ad esempio, nel contesto delle decisioni sulla cauzione, è principalmente la disparità nei tassi di detenzione preventiva dei non recidivi a essere preoccupante, piuttosto che le disparità nei tassi di rilascio anticipato dei recidivi. Sebbene sia vero che il rilascio di potenziali recidivi comporta un costo sotto forma di minaccia alla sicurezza pubblica, tale costo dipende dall'errore totale e non dalla distribuzione di tale errore tra i gruppi. Pertanto, non è necessariamente significativo confrontare semplicemente i tassi di errore tra i gruppi.

Parità del tasso di errore e visione intermedia delle pari opportunità

Ricordiamo che la visione mediana dell'uguaglianza di opportunità tiene conto delle condizioni sociali storiche e attuali che possono influenzare il motivo per cui le qualifiche delle persone possono differire.

⁶Detto questo, ci sono state proposte per immaginare un sistema alternativo in cui il grado di punizione sia calibrato sulla forza delle prove. Schauer, Federico. Profili, probabilità e stereotipi. Cambridge, MA: Harvard University Press, 2006.

Per comprendere un sistema decisionale rispetto alla visione mediana, è fondamentale sapere se gli effetti delle decisioni stesse potrebbero perpetuare queste condizioni nella società.

Sfortunatamente, questo è difficile da fare con i dati disponibili al momento del processo decisionale, soprattutto se le caratteristiche (che codificano le qualifiche dei soggetti decisionali) non sono disponibili. Una cosa che possiamo fare anche senza le funzionalità è osservare le differenze nei tassi di base (cioè i tassi ai quali i diversi gruppi ottengono i risultati desiderati, come il rimborso del prestito o il successo lavorativo). Se i tassi di base sono significativamente diversi – e se assumiamo che le differenze individuali in capacità e ambizione si annullino a livello di gruppo – ciò suggerisce che le qualifiche delle persone possono differire a causa di circostanze indipendenti dall'individuo.

Ma i tassi di base da soli non fanno luce sulla possibilità che il classificatore possa perpetuare le disuguaglianze esistenti. Per questa analisi, ciò che è importante è se il classificatore impone un onere diseguale ai diversi gruppi. Esistono molti modi ragionevoli per misurare l'onere, ma poiché consideriamo particolarmente grave un tipo di errore – classificare erroneamente qualcuno come immeritevole o ad alto rischio – possiamo considerare il tasso di tale errata classificazione tra i membri di un gruppo come un proxy per il peso imposto a quel gruppo. Ciò è particolarmente vero se consideriamo la possibilità di effetti di ricaduta: ad esempio, negare la liberazione anticipata ha effetti sulle famiglie e sulle comunità degli imputati.

Quando un gruppo è gravato da tassi di errore sproporzionalmente elevati, ciò suggerisce che il sistema potrebbe perpetuare cicli di disuguagliaza. In effetti, Aziz Huq sostiene che, per questa ragione, il sistema di giustizia penale rafforza la stratificazione razziale, e questa è la principale disuguaglianza razziale nella giustizia penale algoritmica.¹⁶⁹ Per essere chiari, l'effetto delle istituzioni sulle comunità è una questione empirica e causale che non può essere affrontata ridotto ai tassi di errore, ma dati i limiti dei dati osservativi disponibili nei tipici scenari decisionali, i tassi di errore rappresentano un punto di partenza per indagare su questa questione. Ciò fornisce una chiara ragione per cui i tassi di errore hanno un certo significato morale. Ma si noti che il riscontro di una disparità nel tasso di errore, di per sé, non suggerisce alcun intervento particolare.

Cosa fare in caso di disparità del tasso di errore

Raccogliere più dati e investire nel miglioramento della tecnologia di classificazione è un modo per mitigare potenzialmente la disparità del tasso di errore. Normalmente, diamo una notevole deferenza al decisore sul compromesso tra costo di raccolta dei dati e accuratezza del modello. Questa deferenza, soprattutto nelle applicazioni del settore privato, si basa sull'idea che gli interessi del decisore e dei soggetti decisionali sono generalmente allineati. Ad esempio, lasciamo agli istituti di credito la decisione su quanto accurate dovrebbero essere le loro previsioni. Se, invece, ai finanziatori fosse richiesto di essere molto accurati nelle loro previsioni, potrebbero concedere prestiti solo nei casi più sicuri, privando molte persone della possibilità di ottenere prestiti, oppure potrebbero fare di tutto per raccogliere dati sui mutuatari, aumentando la probabilità di ottenere prestiti. costo di gestione del sistema e scaricando parte di tale costo sui mutuatari.

L'argomentazione di cui sopra considera solo il benessere totale e non il modo in cui i benefici e

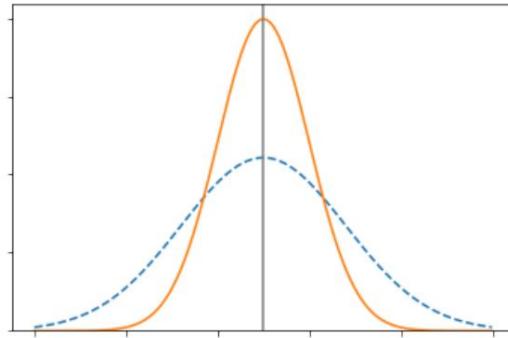


Figura 4.2: Densità di probabilità dei punteggi di rischio per due gruppi e soglia di classificazione. Nelle illustrazioni di questa sezione presupponiamo che il punteggio sia perfettamente calibrato. Il gruppo mostrato con una linea continua ha un tasso di errore più elevato.

Intuitivamente ciò accade perché la massa di probabilità è più concentrata (cioè la funzione di punteggio è peggiore nel distinguere tra i membri di questo gruppo). La raccolta di più dati avvicinerebbe potenzialmente la curva solida alla curva tratteggiata, mitigando la disparità del tasso di errore.

i costi sono distribuiti tra persone e gruppi. Quando introduciamo considerazioni distributive, ci sono molti scenari in cui è giustificabile ridurre la differenza nei confronti dei decisori, e la presenza di disparità nel tasso di errore è uno di questi scenari. In questo caso, richiedere al decisore di mitigare la disparità nel tasso di errore può essere visto come chiedere loro di sostenere parte del costo che viene scaricato su alcuni individui e gruppi.

Sebbene in alcuni casi il miglioramento dell'accuratezza complessiva del classificatore possa colmare la disparità, in altri casi potrebbe lasciarla invariata o addirittura peggiorarla. L'accuratezza è limitata da quella del classificatore ottimale e ricordiamo che il classificatore ottimale non soddisfa necessariamente la parità del tasso di errore. Come esempio concreto, supponiamo che le inadempienze sui prestiti avvengano principalmente a causa della perdita inaspettata di posti di lavoro, che un gruppo di richiedenti il prestito abbia lavori più precari che sono a maggior rischio di licenziamento e che i licenziamenti non siano prevedibili al momento della decisione. In questo scenario, i miglioramenti nella raccolta e nella classificazione dei dati non produrranno la parità del tasso di errore.

Di fronte a questa limitazione intrinseca, potrebbe essere forte la tentazione di eseguire una fase di aggiustamento che raggiunga la parità del tasso di errore, ad esempio diverse soglie di rischio per gruppi diversi. Un modo per farlo sarebbe senza peggiorare la situazione di nessuno rispetto a un classificatore non vincolato. Ad esempio, un finanziatore potrebbe utilizzare una soglia di rischio più indulgente per un gruppo per ridurre il tasso di errore. Ciò violerebbe la visione ristretta delle pari opportunità, poiché persone appartenenti a gruppi diversi con lo stesso punteggio di rischio potrebbero essere trattate in modo diverso. Se l'intervento sia ancora giustificato è una questione normativa difficile a cui manca una risposta uniforme.

In altre situazioni, anche questo potrebbe non essere possibile. Ad esempio, l'intervento può aumentare il rischio del creditore a tal punto da farlo fallire.

Infatti, se i tassi di base sono così diversi da farci prevedere grandi disparità nei tassi di errore che non possono essere mitigate da interventi come la raccolta dei dati, allora ciò suggerisce

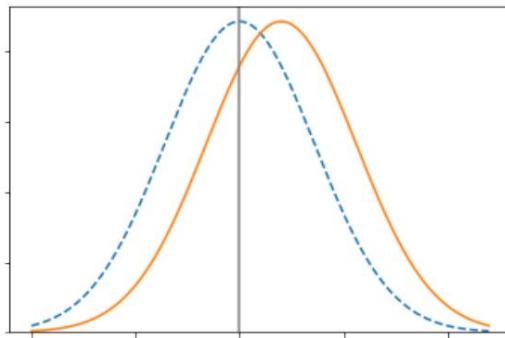


Figura 4.3: Densità di probabilità dei punteggi di rischio per due gruppi e soglia di classificazione. Anche in questo caso il gruppo solido ha un tasso di errore più elevato, in particolare un tasso di falsi positivi più elevato, dove i falsi positivi sono persone erroneamente classificate come ad alto rischio. Ma questa volta è perché il gruppo solido ha un tasso base più elevato (la curva è spostata a destra rispetto al gruppo tratteggiato). È improbabile che la raccolta di più dati riduca la disparità del tasso di errore.

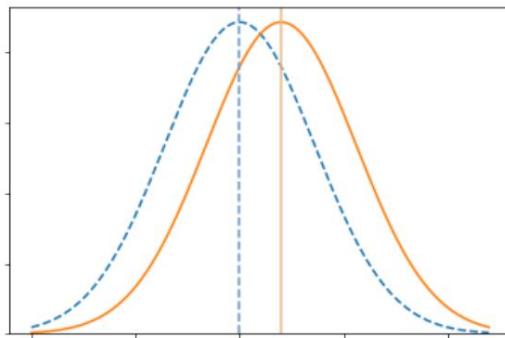


Figura 4.4: Densità di probabilità dei punteggi di rischio per due gruppi e due diverse soglie di classificazione che danno luogo a tassi di errore uguali.

che l'uso del processo decisionale predittivo è di per sé problematico, e forse dovremmo eliminare il sistema o applicare interventi più fondamentali.

In sintesi, la parità del tasso di errore non ha una relazione diretta con alcun singolo principio normativo. Ma coglie qualcosa sia nella visione ristretta che in quella mediana dell'uguaglianza di opportunità. È anche un modo per incentivare i decisori a investire nell'equità e a mettere in discussione l'adeguatezza del processo decisionale predittivo.

Alternative per realizzare la visione mediana delle pari opportunità

Abbiamo discusso di come la parità del tasso di errore abbia qualche relazione con la visione mediana delle pari opportunità. Ma ci sono molti altri possibili interventi che i decisori potrebbero adottare per cercare di realizzare la visione mediana delle pari opportunità che non corrispondono a nessuno dei criteri discussi nel capitolo 3. La visione mediana è una nozione intrinsecamente confusa, che lascia molti dubbi. spazio per decidere fino a che punto vogliamo sottovalutare le differenze apparenti delle persone e il modo in cui farlo. Ecco alcuni altri modi in cui possiamo provare a renderlo operativo . Non sorprende che tutto ciò viola la visione ristretta delle pari opportunità.

I decisori potrebbero riconsiderare gli obiettivi che stanno perseguiti in modo tale che il processo decisionale che cerca di raggiungere questi obiettivi generi risultati meno disparati. Ad esempio, i datori di lavoro potrebbero scegliere una variabile target diversa che viene percepita come un indicatore altrettanto valido per il loro obiettivo, ma la cui previsione accurata porta a una disparità meno significativa nei risultati per i diversi gruppi.¹⁷⁰

Potrebbero esplorare se sia possibile addestrare modelli alternativi con un grado di accuratezza simile a quello del modello originale, ma che produca disparità minori nella velocità con cui i membri di gruppi diversi raggiungono il risultato desiderato o sono soggetti a valutazioni errate.¹⁷¹ Empiricamente, ciò sembra essere possibile in molti casi, anche per processi decisionali ad alto rischio.¹⁷² Potrebbero sacrificare

una buona dose di accuratezza apparente nella convinzione che vi sia un grave errore di misurazione e che le persone di alcuni gruppi siano in realtà molto più qualificate di loro. potrebbero apparire (assumiamo che non sia possibile correggere esplicitamente l'errore di misurazione e che l'appartenenza al gruppo non sia codificata in modo sufficientemente ridondante nelle caratteristiche, impedendo al classificatore ottimale di tenere conto automaticamente dell'errore di

misurazione).¹⁷³ Infine, potrebbero rinunciare ad alcune delle benefici che avrebbero potuto ottenere nell'ambito del processo decisionale originale in modo da fornire importanti benefici ai gruppi che sono stati soggetti a maltrattamenti in passato. Per fare ciò, potrebbero trattare i membri di determinati gruppi in modo controfattuale, come se non avessero sperimentato l' ingiustizia che li rende meno qualificati al momento del processo decisio-

Riepilogo

L'equità è spesso concettualizzata come uguaglianza di opportunità. Ma in questo capitolo abbiamo visto che esistono diversi modi per comprendere l'uguaglianza di opportunità. Le differenze tra loro sono al centro del motivo per cui l'equità è così contestata

argomento. Tutti e tre i punti di vista possono essere visti nei dibattiti politici contemporanei. Questa visione ristretta è in linea con ciò che spesso si intende con il termine meritocrazia. La visione centrale guida gli sforzi in materia di diversità, equità e inclusione (DEI) in molti luoghi di lavoro. La visione ampia è troppo ampia per trovare molto sostegno per un'implementazione completa, ma le idee alla base di essa emergono nei dibattiti su argomenti come le riparazioni.¹⁷⁴

I punti di vista differiscono lungo molti assi, compreso ciò che cercano di ottenere; come comprendono le cause delle attuali differenze tra i gruppi (e se cercano di capirle del tutto); e come distribuire il costo del miglioramento dei gruppi storicamente svantaggiati.

Tabella 4.2: Opinioni sulle pari opportunità e loro affini formali

	Obiettivo	Criteri formali correlati
Stretto visualizzazione	Garantire che le persone che hanno le stesse qualifiche per un'opportunità abbiano le stesse possibilità di ottenerla	Criterio di similarità, equità meritocratica, calibrazione per gruppo
Vista centrale	Differenze di sconto dovute a ingiustizie passate che spiegano le attuali differenze nelle qualifiche	Criterio di somiglianza, Uguaglianza formale di opportunità di Roemer, parità del tasso di errore
Visione ampia	Garantire che persone con pari capacità e ambizione siano in grado di realizzare il proprio potenziale altrettanto bene	Parità demografica

Nell'ultima parte del capitolo, abbiamo tentato di collegare queste nozioni morali ai criteri statistici del Capitolo 3. Attraverso questo esercizio sono emerse connessioni vaghe, ma, in definitiva, nessuno dei criteri statistici è fortemente ancorato a fondamenti normativi.

Ma anche queste somiglianze approssimative illustrano un punto importante sui risultati di impossibilità del Capitolo 3. I risultati di impossibilità non sono una sorta di artefatto del processo decisionale statistico; rivelano semplicemente dilemmi morali. Una volta riconosciute le difficoltà morali sottostanti, queste tensioni matematiche sembrano molto meno sorprendenti.

Ad esempio, un approccio che fa previsioni accurate basate sugli attributi attualmente osservabili delle persone, e poi prende decisioni basate su tali previsioni (calibrazione) non si tradurrà in uguaglianza di risultati (indipendenza) fintanto che gruppi diversi hanno in media qualifiche diverse. . Allo stesso modo, i suoi risultati differiscono anche da un approccio che è disposto a trattare persone apparentemente simili in modo diverso nel tentativo di egualizzare l'onere gravante su gruppi diversi (parità del tasso di errore). Gli approcci differiscono anche nella misura in cui gli errori di misurazione sono visti come responsabilità del decisore e in chi dovrebbe sostenere i costi degli interventi di equità.

Uno dei motivi per cui i fondamenti normativi dei criteri di equità statistica sono instabili è che l'indipendenza condizionale non ci fornisce un vocabolario su cui ragionare.

le cause delle disparità tra i gruppi o gli effetti degli interventi. Cercheremo di affrontare queste limitazioni nel prossimo capitolo.

5

Causalità

Il nostro punto di partenza è la differenza tra un'osservazione e un'azione. Ciò che vediamo nell'osservazione passiva è il modo in cui gli individui seguono il loro comportamento di routine, le loro abitudini e le loro inclinazioni naturali. L'osservazione passiva riflette lo stato del mondo proiettato su un insieme di caratteristiche che abbiamo scelto di evidenziare. I dati che raccogliamo dall'osservazione passiva mostrano un'istantanea del nostro mondo così com'è.

Ci sono molte domande a cui possiamo rispondere solo con l'osservazione passiva: i conducenti di 16 anni hanno un tasso di incidenza di incidenti stradali più elevato rispetto ai guidatori di 18 anni? Formalmente, la risposta corrisponde a una differenza di probabilità condizionate assumendo di modellare la popolazione come una distribuzione come abbiamo fatto nel capitolo precedente. Possiamo calcolare la probabilità condizionata di un incidente stradale dato che l'età del conducente è 16 anni e sottrarre da essa la probabilità condizionata di un incidente stradale dato che l'età è 18 anni. Entrambe le probabilità condizionali possono essere stimate da un campione sufficientemente ampio estratto dalla distribuzione, presupponendo che vi siano conducenti sia di 16 che di 18 anni. La risposta alla domanda che ci siamo posti rientra saldamente nell'ambito delle statistiche osservative.

Ma le domande importanti spesso non sono di natura osservativa. Le vittime della strada diminuirebbero se aumentassimo di due anni l'età legale per guidare? Sebbene la domanda sembri simile in superficie, ci rendiamo subito conto che richiede una visione fondamentalmente diversa. Piuttosto che chiedere la frequenza di un evento nel nostro mondo manifesto, questa domanda chiede l'effetto di un'azione ipotetica.

Di conseguenza, la risposta non è così semplice. Anche se i conducenti più anziani hanno un tasso di incidenza di incidenti stradali più basso, ciò potrebbe semplicemente essere una conseguenza di una maggiore esperienza di guida. Non vi è alcuna ragione ovvia per cui un diciottenne con due mesi di viaggio avrebbe meno probabilità di essere coinvolto in un incidente rispetto, ad esempio, a un sedicenne con la stessa esperienza. Possiamo provare ad affrontare questo problema mantenendo fisso il numero di mesi di esperienza di guida, confrontando individui di età diverse. Ma ci imbattiamo rapidamente nelle sottigliezze. E se i diciannovenne con due mesi di esperienza alla guida corrispondessero a individui eccezionalmente prudenti e quindi – per loro naturale inclinazione – non solo guidano meno, ma anche più prudentemente? Cosa accadrebbe se tali individui vivessero prevalentemente in regioni in cui le condizioni del traffico differiscono in modo significativo da quelle in aree in cui le persone sentono un maggiore bisogno di guidare in giovane età?

Possiamo pensare a numerose altre strategie per rispondere alla domanda iniziale

se l'innalzamento dell'età legale per guidare riduce gli incidenti stradali. Potremmo confrontare paesi con diverse età legali di guida, ad esempio gli Stati Uniti e la Germania.

Ma ancora una volta, questi paesi differiscono in molti altri aspetti potenzialmente rilevanti, come ad esempio l'età legale per bere alcolici.

Inizialmente, il ragionamento causale è un quadro concettuale e tecnico per affrontare domande sull'effetto di ipotetiche azioni o interventi. Una volta compreso qual è l'effetto di un'azione, possiamo ribaltare la questione e chiederci quale azione abbia plausibilmente causato un evento. Questo ci dà un linguaggio formale per parlare di causa ed effetto.

Non tutte le domande sulla causa sono ugualmente facili da affrontare. Alcune domande sono eccessivamente generiche, come: "Qual è la causa del successo?" Altre domande sono troppo specifiche: "Che cosa ha causato il tuo interesse per la filosofia tedesca del XIX secolo?" Nessuna delle due domande potrebbe avere una risposta chiara. L'inferenza causale ci fornisce un linguaggio formale per porre queste domande, in linea di principio, ma non rende facile scegliere le domande giuste. Né banalizza il compito di trovare e interpretare la risposta a una domanda. Soprattutto nel contesto dell'equità, la difficoltà sta spesso nel decidere quale sia la domanda a cui l'inferenza causale è la risposta.

In questo capitolo svilupperemo una comprensione tecnica della causalità sufficiente a supportare almeno tre scopi diversi. Il primo è concettualizzare e affrontare alcuni limiti delle tecniche di osservazione che abbiamo visto nel capitolo 3. Il secondo è fornire strumenti che aiutino nella progettazione di interventi che raggiungano in modo affidabile l'effetto desiderato. Il terzo è impegnarsi nell'importante dibattito normativo su quando e in che misura il ragionamento sulla discriminazione e sull'equità richiede una comprensione causale.

I limiti dell'osservazione

Prima di sviluppare qualsiasi nuovo formalismo, è importante capire perché ne abbiamo bisogno. Per capire perché ci rivolgiamo al venerabile esempio dell'ammissione dei laureati all'Università della California, Berkeley nel 1973.

175 I dati storici mostrano che

12763 candidati sono stati presi in considerazione per l'ammissione a uno dei 101 dipartimenti e specializzazioni interdipartimentali. Delle 4.321 donne che hanno presentato domanda sono state ammesse circa il 35%, degli 8.442 uomini che hanno presentato domanda sono stati ammessi il 44%.

I test standard di significatività statistica suggeriscono che sarebbe altamente improbabile che la differenza osservata fosse il risultato della fluttuazione del campione se non ci fossero differenze nei tassi di accettazione sottostanti.

Un modello simile esiste se guardiamo alle decisioni di ammissione aggregate dei sei dipartimenti più grandi. Il tasso di accettazione in tutti e sei i dipartimenti per gli uomini è di circa il 44%, mentre per le donne è solo del 30% circa, ancora una volta una differenza significativa.

Riconoscendo che i dipartimenti hanno autonomia su chi ammettere, possiamo esaminare i pregiudizi di genere di ciascun dipartimento.

Tabella 5.1: dati di ammissione all'UC Berkeley dal 1973.

	Uomini	Donne	
Dipartimento Applicato Ammesso (%) Applicato Ammesso (%)			
A	825 520 325 417	62	108
B		60	25
C		37	593
D		33	375
E	191	28	393
F	373	6	341
			82
			68
			34
			35
			24
			7

Ciò che possiamo vedere dalla tabella è che quattro dei sei dipartimenti più grandi lo mostrano un tasso di accettazione più elevato tra le donne, mentre due mostrano un tasso di accettazione più elevato per gli uomini. Tuttavia, questi due dipartimenti non possono spiegare la grande differenza nei tassi di accettazione che abbiamo osservato in totale. Quindi, sembra che maggiore sia il tasso di accettazione per gli uomini che abbiamo osservato nel complesso sembra essersi invertito il livello di dipartimento.

Tali inversioni sono talvolta chiamate il paradosso di Simpson, anche se matematicamente non costituiscono una sorpresa. È un fatto di probabilità condizionata che possa esserci un evento Y (qui, accettazione), un attributo A (qui, genere femminile considerato binario variabile casuale) e una variabile casuale Z (qui, scelta del dipartimento) tale che:

1. $P\{Y | A\} < P\{Y | \neg A\}$
2. $P\{Y | A, Z = z\} > P\{Y | \neg A, Z = z\}$ per tutti i valori z che sono casuali presuppone la variabile Z.

Il paradosso di Simpson provoca tuttavia disagio ad alcuni, a causa dell'intuizione suggerisce che una tendenza che vale per tutte le sottopopolazioni dovrebbe valere anche per livello di popolazione.

Il motivo per cui il paradosso di Simpson è rilevante per la nostra discussione è che lo è una conseguenza di come tendiamo a interpretare erroneamente le informazioni condizionali codificare le probabilità. Ricordiamo che corrisponde un'affermazione di probabilità condizionata all'osservazione passiva. Ciò che vediamo qui è un'istantanea del comportamento normale di donne e uomini che si iscrivono alla scuola di specializzazione presso l'UC Berkeley nel 1973.

Ciò che risulta evidente dai dati è che il genere influenza la scelta del dipartimento. Le donne e gli uomini sembrano avere preferenze diverse per i diversi campi di studio. Inoltre, diversi dipartimenti hanno criteri di ammissione diversi. Alcuni hanno valori inferiori tassi di accettazione, alcuni più alti. Pertanto, una spiegazione per i dati che vediamo è che le donne hanno scelto di candidarsi a dipartimenti più competitivi, venendo quindi respinte ad un tasso più elevato rispetto agli uomini.

In effetti, questa è la conclusione tratta dallo studio originale:

La distorsione nei dati aggregati non deriva da alcun modello di discriminazione da parte dei comitati di ammissione, il che nel complesso sembra abbastanza giusto, ma evidentemente da uno screening preliminare ai livelli precedenti del sistema educativo.

Le donne vengono indirizzate, a causa della loro socializzazione e della loro istruzione, verso campi di studi universitari che sono generalmente più affollati, meno produttivi di titoli di studio completati e meno ben finanziati, e che spesso offrono prospettive di impiego professionale più povere.¹⁷⁵

In altre parole, l'articolo concludeva che la fonte dei pregiudizi di genere nelle ammissioni era un problema di pipeline: senza alcun illecito da parte del comitato di ammissione, le donne venivano "sbarrate dalla loro socializzazione" avvenuta in una fase precedente della loro vita.

È difficile discutere questa conclusione sulla base dei soli dati disponibili.

La questione della discriminazione, tuttavia, è lungi dall'essere risolta. Possiamo chiederci innanzitutto perché le donne si sono candidate a dipartimenti più competitivi. Ci sono diverse possibili ragioni. Forse i dipartimenti meno competitivi, come le scuole di ingegneria, all'epoca non erano accoglienti nei confronti delle donne. Questo potrebbe essere stato un modello generale all'epoca o specifico per l'università. Forse alcuni dipartimenti avevano precedenti di cattivo trattamento delle donne che erano noti ai richiedenti. Forse il dipartimento ha pubblicizzato il programma in un modo che ha scoraggiato le donne dal fare domanda.

Inoltre, i dati in nostro possesso non mostrano alcuna misurazione della qualificazione di un richiedente. È possibile che, a causa dell'autoselezione, le donne che si iscrivevano alle scuole di ingegneria nel 1973 fossero troppo qualificate rispetto ai loro coetanei. In questo caso, un tasso di accettazione paritario tra uomini e donne potrebbe effettivamente essere un segno di discriminazione.

Non c'è modo di sapere cosa sia successo dai dati in nostro possesso. Esistono molteplici scenari possibili con diverse interpretazioni e conseguenze che non possiamo distinguere dai dati a nostra disposizione. A questo punto, abbiamo due scelte.

Uno è progettare un nuovo studio e raccogliere più dati in un modo che possa portare a un risultato più conclusivo. L'altro è discutere quale scenario sia più probabile in base alle nostre convinzioni e ipotesi plausibili sul mondo. L'inferenza causale è utile in entrambi i casi. Da un lato, può essere utilizzato come guida nella progettazione di nuovi studi. Può aiutarci a scegliere quali variabili includere, quali escludere e quali mantenere costanti. D'altro canto, i modelli causali possono servire come meccanismo per incorporare la conoscenza del dominio scientifico e scambiare ipotesi plausibili con conclusioni plausibili.

Modelli causali

Svilupperemo concetti formali quanto basta per affrontare il dibattito tecnico e normativo sulla causalità e la discriminazione. L'argomento è molto più profondo di quello che possiamo esplorare in questo capitolo.

Sceglieremo i modelli causali strutturali come base della nostra discussione formale poiché hanno il vantaggio di fornire una solida base per le varie nozioni causali che incontreremo. Il modo più semplice per concettualizzare un modello causale strutturale è come un programma per generare una distribuzione da variabili di rumore indipendenti attraverso una sequenza di istruzioni formali. Analizziamo questa affermazione. Immagina invece di campioni da una distribuzione, qualcuno ti ha dato un programma per computer passo dopo passo

per generare campioni in autonomia partendo da un seme casuale. Il processo non è diverso da come scriveresti il codice. Inizi da un semplice seme casuale e costruisci costrutti sempre più complessi. Questo è fondamentalmente ciò che è un modello causale strutturale , tranne per il fatto che ogni compito utilizza il linguaggio della matematica piuttosto che qualsiasi sintassi di programmazione concreta.

Un primo esempio

Cominciamo con un esempio di giocattolo non destinato a catturare il mondo reale. Immagina una popolazione ipotetica in cui un individuo si esercita regolarmente con probabilità $1/2$. Con probabilità $1/3$, l'individuo ha una predisposizione latente a sviluppare sovrappeso che si manifesta in assenza di esercizio fisico regolare. Allo stesso modo, in assenza di esercizio fisico, le malattie cardiache si verificano con probabilità $1/3$. Indichiamo con X la variabile indicatore dell'esercizio fisico regolare, con W quella del peso eccessivo e con H l'indicatore delle malattie cardiache. Di seguito è riportato un modello causale strutturale per generare campioni da questa ipotetica popolazione. Per facilitare la descrizione, indichiamo con $B(p)$ una variabile casuale di Bernoulli con bias p , ovvero un lancio della moneta distorto che assume valore 1 con probabilità p e valore 0 con probabilità $1 - p$.

1. Variabili casuali di Bernoulli indipendenti campione $U_1 \sim B(1/2)$, $U_2 \sim B(1/3)$, $U_3 \sim B(1/3)$.
2. $X := U_1$
3. $W :=$ se $X = 1$ allora 0 altrimenti U_2
4. $H :=$ se $X = 1$ allora 0 altrimenti U_3

Confrontate questa descrizione generativa della popolazione con un campione casuale estratto dalla popolazione. Dalla descrizione del programma si vede subito che nella nostra ipotetica popolazione l'esercizio fisico previene sia il sovrappeso che le malattie cardiache, ma in assenza di esercizio i due sono indipendenti. All'inizio, il nostro programma genera una distribuzione congiunta sulle variabili casuali (X, W, H). Possiamo calcolare le probabilità con questa distribuzione. Ad esempio, la probabilità di malattie cardiache secondo la distribuzione specificata dal nostro modello è $1/2 \cdot 1/3 = 1/6$. Possiamo anche calcolare la probabilità condizionata di malattie cardiache in caso di sovrappeso. Dall'evento $W = 1$ possiamo dedurre che l'individuo non fa attività fisica, quindi la probabilità di malattie cardiache in caso di sovrappeso aumenta a $1/3$ rispetto al basale di $1/6$.

Ciò significa che nel nostro modello il sovrappeso provoca malattie cardiache? La risposta è no in quanto è intuitiva visto il programma per generare la distribuzione. Ma vediamo come discuteremo formalmente questo punto. Avere un programma per generare una distribuzione è sostanzialmente più potente che avere semplicemente l'accesso al campionamento. Una ragione è che possiamo manipolare il programma nel modo che vogliamo, presupponendo che ci ritroveremo comunque con un programma valido. Potremmo, ad esempio, impostare $W := 1$, ottenendo una nuova distribuzione. Il programma risultante è simile al seguente:

2. $X := U_1$
3. $W := 1$

4. $H := \text{se } X = 1 \text{ allora } 0 \text{ altrimenti } U_3$

Questo nuovo programma specifica una nuova distribuzione. Possiamo nuovamente calcolare la probabilità di malattie cardiache con questa nuova distribuzione. Otteniamo ancora $1/6$. Questo semplice calcolo rivela un'intuizione significativa. La sostituzione $W := 1$ non corrisponde ad un condizionamento su $W = 1$. Una è un'azione, anche se in questo caso irrilevante. L'altra è un'osservazione da cui possiamo trarre delle deduzioni. Se osserviamo che un individuo è in sovrappeso, possiamo dedurre che ha un rischio maggiore di malattie cardiache (nel nostro esempio del giocattolo). Tuttavia, ciò non significa che la riduzione del peso corporeo eviterebbe le malattie cardiache. Non lo sarebbe nel nostro esempio. La sostituzione attiva $W := 1$ crea invece una nuova ipotetica popolazione in cui tutti gli individui sono in sovrappeso con tutto ciò che ciò comporta nel nostro modello.

Approfondiamo ancora un po' questo punto considerando un'altra ipotetica popolazione. L'azione, specificata dalle equazioni:

- 2. $W := U_2$
- 3. $X := \text{se } W = 0 \text{ allora } 0 \text{ altrimenti}$
- U1 4. $H := \text{se } X = 1 \text{ allora } 0 \text{ altrimenti } U_3$

In questa popolazione le abitudini di esercizio sono determinate dal peso corporeo. Le persone in sovrappeso scelgono di fare esercizio con una certa probabilità, ma questa è l'unica ragione per cui qualcuno dovrebbe esercitarsi. Le malattie cardiache si sviluppano in assenza di esercizio fisico. La sostituzione $W := 1$ in questo modello porta ad una maggiore probabilità di esercizio fisico, riducendo quindi la probabilità di malattie cardiache. In questo caso il condizionamento su $W = 1$ ha lo stesso effetto. Entrambi portano ad una probabilità di $1/6$.

Ciò che vediamo è che fissare una variabile mediante sostituzione può o meno corrispondere a una probabilità condizionata. Questa è una resa formale del nostro punto precedente secondo cui l'osservazione non è azione. Una sostituzione corrisponde a un'azione che eseguiamo. Sostituendo un valore interrompiamo il corso naturale dell'azione catturato dal nostro modello. Questo è il motivo per cui l'operazione di sostituzione è talvolta chiamata operatore do, scritto come $\text{do}(W := 1)$.

I modelli causali strutturali ci forniscono un calcolo formale per ragionare sull'effetto di azioni ipotetiche. Vedremo come ciò crei una base formale per tutte le diverse nozioni causali che incontreremo in questo capitolo.

Modelli causali strutturali, più formalmente

Formalmente, un modello causale strutturale è una sequenza di compiti per generare una distribuzione congiunta a partire da variabili di rumore indipendenti. Eseguendo la sequenza di assegnamenti costruiamo in modo incrementale un insieme di variabili casuali distribuite congiuntamente. Un modello causale strutturale quindi non solo fornisce una distribuzione congiunta, ma anche una descrizione di come la distribuzione congiunta può essere generata dalle variabili elementari del rumore. La definizione formale è un po' macchinosa rispetto alla nozione intuitiva.

Definizione 4. Un modello causale strutturale M è dato da un insieme di variabili X_1, \dots, X_d e corrispondenti assegnazioni della forma

$$X_i := f_i(P_i, U_i), \quad i = 1, \dots, d.$$

Qui, $P_i \subseteq \{X_1, \dots, X_d\}$ è un sottoinsieme delle variabili che chiamiamo genitori di X_i . Le variabili casuali U_1, \dots, U_d sono chiamate variabili di rumore, che richiediamo siano congiuntamente indipendenti. Il grafo causale corrispondente al modello causale strutturale è il grafo diretto che ha un nodo per ogni variabile X_i con archi entranti da tutti i genitori P_i .

Esaminiamo i concetti formali introdotti in questa definizione in modo un po' più dettagliato. Le variabili di rumore che compaiono nella definizione modellano i fattori esogeni che influenzano il sistema. Considera, ad esempio, come il tempo influenza il ritardo su un percorso di traffico che scegli. A causa della difficoltà di modellare l'influenza del tempo in modo più preciso, potremmo considerare il ritardo indotto dal tempo come un fattore esogeno che entra nel modello come variabile di rumore. La scelta delle variabili esogene e la loro distribuzione può avere conseguenze importanti sulle conclusioni che traiamo da un modello.

I nodi genitori P_i del nodo i in un modello causale strutturale sono spesso chiamati le cause dirette di X_i . Allo stesso modo, chiamiamo X_i l'effetto diretto delle sue cause dirette P_i . Ricordiamo la nostra ipotetica popolazione in cui l'aumento di peso è stato determinato dalla mancanza di esercizio fisico tramite l'assegnazione $W := \min\{U_1, 1 - X\}$. Qui diremmo che l'esercizio fisico (o la sua mancanza) è una causa diretta dell'aumento di peso.

Il modello causale strutturale è una raccolta di ipotesi formali su come interagiscono determinate variabili. Ciascuna assegnazione specifica una funzione di risposta. Possiamo pensare ai nodi come se ricevessero messaggi dai loro genitori e agissero in base a questi messaggi, nonché all'influenza di una variabile di rumore esogena.

In che misura un modello causale strutturale è conforme alla realtà è una questione separata e difficile su cui torneremo più dettagliatamente in seguito. Per ora, pensiamo a un modello causale strutturale come se formalizzasse ed esponesse una serie di ipotesi su un processo di generazione di dati. Pertanto, modelli diversi possono esporre diversi scenari ipotetici e servire come base per la discussione. Quando facciamo affermazioni su causa ed effetto in riferimento a un modello, non intendiamo suggerire che queste relazioni valgano necessariamente nel mondo reale. Se lo faranno dipende dalla portata, dallo scopo e dalla validità del nostro modello, che potrebbe essere difficile da dimostrare.

Non è difficile dimostrare che un modello causale strutturale definisce un'unica distribuzione congiunta sulle variabili (X_1, \dots, X_d) tale che $X_i = f_i(P_i, U_i)$. È conveniente introdurre una nozione di probabilità sotto questa distribuzione. Quando M denota un modello causale strutturale, scriveremo la probabilità di un evento E sotto la distribuzione congiunta implicata come $PM(E)$. Per acquisire familiarità con la notazione, indichiamo con M il modello causale strutturale per la popolazione ipotetica in cui sia l'aumento di peso che le malattie cardiache sono direttamente causati dall'assenza di esercizio fisico. Abbiamo calcolato in precedenza che la probabilità di malattia cardiaca in questo modello è $PM(H) = 1/6$.

In quanto segue deriveremo da questa singola definizione di modello causale strutturale tutte le diverse nozioni e terminologie di cui avremo bisogno in questo capitolo.

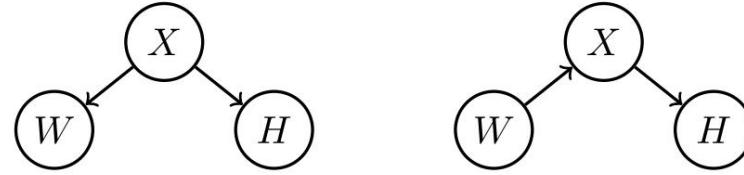


Figura 5.1: Diagrammi causali per gli esempi di malattie cardiache.

Nel complesso, limiteremo la nostra attenzione alle assegnazioni acicliche. Molti sistemi del mondo reale sono naturalmente descritti come sistemi dinamici con stato con cicli di feedback chiusi. Esistono alcuni modi per gestire tali sistemi a circuito chiuso. Ad esempio, spesso i cicli possono essere spezzati introducendo variabili dipendenti dal tempo, come ad esempio gli investimenti al tempo 0 fanno crescere l'economia al tempo 1 che a sua volta fa crescere gli investimenti al tempo 2, continuando così fino a un orizzonte temporale t scelto. Questa elaborazione è chiamata srotolamento di un sistema dinamico.

Grafici causali

Abbiamo visto come i modelli causali strutturali danno origine naturalmente a grafici causali che rappresentano graficamente la struttura di assegnazione del modello. Possiamo andare anche nella direzione opposta, semplicemente considerando i grafici orientati come segnaposto per un modello causale strutturale non specificato che ha la struttura di assegnazione data dal grafico. I grafici causali sono spesso chiamati diagrammi causali. Useremo questi termini in modo intercambiabile.

I grafici causali per le due ipotetiche popolazioni del nostro esempio di malattia cardiaca hanno ciascuno due bordi e gli stessi tre nodi. Concordano sul legame tra esercizio fisico e malattie cardiache, ma differiscono nella direzione del legame tra esercizio fisico e aumento di peso.

I grafici causali sono convenienti quando le assegnazioni esatte in un modello causale strutturale sono di secondaria importanza, ma ciò che conta sono i percorsi presenti e assenti nel grafico. I grafici ci consentono anche di importare il linguaggio consolidato della teoria dei grafi per discutere le nozioni causali. Possiamo dire, ad esempio, che una causa indiretta di un nodo è un qualsiasi antenato del nodo in un dato grafo causale. In particolare, i grafi causali permettono di distinguere causa ed effetto a seconda che un nodo sia antenato o discendente di un altro nodo.

Diamo un primo sguardo ad alcune importanti strutture grafiche.

Forchette

Una forchetta è un nodo Z in un grafico che ha bordi in uscita verso altre due variabili X e Y . In altre parole, il nodo Z è una causa comune di X e Y . Abbiamo già visto un esempio di forchetta nel nostro esempio di peso ed esercizio : $W \rightarrow X \rightarrow H$. Qui, l'esercizio X influenza sia il peso che le malattie cardiache. Dall'esempio abbiamo anche imparato che Z ha un effetto confondente: ignorando l'esercizio X , abbiamo visto che W e H sembrano essere correlati positivamente. Tuttavia, la correlazione è un semplice risultato di

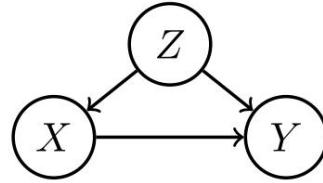


Figura 5.2: Esempio di fork (confondente).

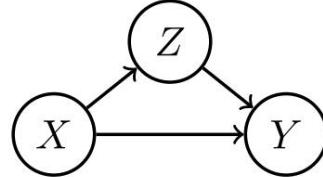


Figura 5.3: Esempio di catena (mediatore).

confondente. Una volta che manteniamo costanti i livelli di esercizio fisico (tramite l'operazione `do`), nel nostro esempio il peso non ha alcun effetto sulle malattie cardiache.

Il confondimento porta a un disaccordo tra il calcolo delle probabilità condizionali (osservazione) e gli interventi (azioni). Gli esempi reali di confusione rappresentano una minaccia comune alla validità delle conclusioni tratte dai dati. Ad esempio, in un noto studio medico un sospetto effetto benefico della terapia ormonale sostitutiva nel ridurre le malattie cardiovascolari è scomparso dopo aver identificato lo stato socioeconomico come variabile confondente.¹⁷⁶

Mediatori

Il caso di una biforcazione è abbastanza diverso dalla situazione in cui Z si trova su un percorso diretto da X a Y . In questo caso, il percorso $X \rightarrow Z \rightarrow Y$ contribuisce all'effetto totale di X su Y . È un percorso causale e quindi uno dei modi in cui X influenza causalmente Y . Ecco perché Z non è un confondente. Chiamiamo invece Z un mediatore.

Abbiamo visto un esempio plausibile di mediatore nel nostro esempio di ammissione alla UC Berkeley. In un grafico causale plausibile, la scelta del dipartimento media le influenze del genere sulla decisione di ammissione. La nozione di mediatore è particolarmente rilevante per il tema dell'analisi della discriminazione, poiché i mediatori possono essere interpretati come il meccanismo alla base di un nesso causale.

Collider

Consideriamo infine un'altra situazione comune: il caso di un collisore. Gli addetti ai conflitti non creano confusione. Infatti, nel grafico sopra, X e Y non sono confusi, il che significa che possiamo sostituire le affermazioni `do` con probabilità condizionali. Tuttavia, accade qualcosa di interessante quando si condiziona su un collisore. La fase di condizionamento può creare una correlazione tra X e Y , un fenomeno chiamato spiegazione.

Un buon esempio dell'effetto di spiegazione, o bias del collisore, è dovuto a Berkson.

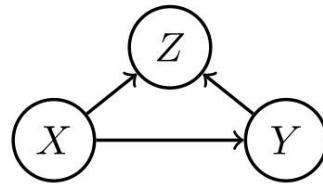


Figura 5.4: Esempio di collisore.

Due malattie indipendenti possono diventare correlate negativamente quando si analizzano i pazienti ospedalizzati. La ragione è che quando una delle due malattie (X o Y) è sufficiente per il ricovero in ospedale (indicato dalla variabile Z), osservare che un paziente ha una malattia rende statisticamente meno probabile l'altra.¹⁷⁷

La legge di Berkson è un ammonimento per l'analisi statistica quando studiamo una cohorte che è stata sottoposta a una regola di selezione. Ad esempio, è in corso un dibattito sull'efficacia dei punteggi GRE nell'istruzione superiore. Alcuni studi^{178,179} sostengono che i punteggi GRE non sono predittivi dei vari esiti di successo in una popolazione di studenti laureati. Tuttavia, occorre prestare attenzione quando si studia l'efficacia dei test didattici, come il GRE, esaminando un campione di studenti ammessi. Dopotutto, gli studenti venivano ammessi in parte sulla base del punteggio del test. È la regola di selezione che introduce il potenziale di bias del collisore.

Interventi ed effetti causali

I modelli causali strutturali ci danno un modo per formalizzare l'effetto di ipotetiche azioni o interventi sulla popolazione entro i presupposti del nostro modello.

Come abbiamo visto prima, tutto ciò di cui avevamo bisogno era la capacità di effettuare sostituzioni.

Sostituzioni e operatore do

Dato un modello causale strutturale M possiamo accettare qualsiasi assegnazione della forma

$$X := f(P, U)$$

e sostituirlo con un altro compito. La sostituzione più comune è con assegna a X un valore costante x:

$$X := x$$

Indicheremo il modello risultante con $M = M[X := x]$ per indicare l'intervento chirurgico che abbiamo eseguito sul modello originale M. In base a questa assegnazione manteniamo X costante rimuovendo l'influenza dei suoi nodi principali e quindi qualsiasi altra variabile nel modello.

Graficamente, l'operazione corrisponde all'eliminazione di tutti gli archi in ingresso al nodo X. I figli di X nel grafico ora ricevono un messaggio fisso x da X quando interrogano il valore del nodo. L'operatore di assegnazione è anche chiamato

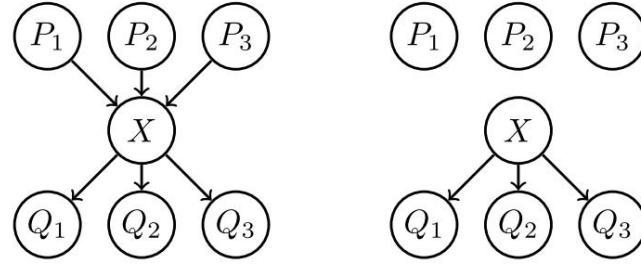


Figura 5.5: Grafico prima e dopo la sostituzione.

do-operator per sottolineare che corrisponde all'esecuzione di un'azione o di un intervento. Abbiamo già la notazione per calcolare le probabilità dopo aver applicato l'operatore do, vale a dire $PM[X:=x]$ (E). Un'altra notazione è popolare e comune: $P\{E$

$$| \text{ fare}(X := x)\} = PM[X:=x] (E)$$

Questa notazione analogizza l'operazione do con la consueta notazione per le probabilità condizionali, ed è spesso conveniente quando si eseguono calcoli che coinvolgono l'operatore do. Tieni presente, tuttavia, che l'operatore do (azione) è fondamentalmente diverso dall'operatore condizionamento (osservazione).

Effetti causali

L'effetto causale di un'azione $X := x$ su una variabile Y si riferisce alla distribuzione della variabile Y nel modello $M[X := x]$. Quando parliamo di effetto causale di una variabile X su un'altra variabile Y ci riferiamo a tutti i modi in cui ponendo X a qualsiasi possibile valore x influenza la distribuzione di Y.

Spesso pensiamo a X come a una variabile di trattamento binaria e siamo interessati a una quantità come

$$EM[X=1] [Y] \circ EM[X=0] [Y].$$

Questa quantità è chiamata effetto medio del trattamento. Ci dice di quanto il trattamento (azione $X := 1$) aumenta l'aspettativa di Y rispetto a nessun trattamento (azione $X := 0$). Gli effetti causali sono quantità di popolazione. Si riferiscono agli effetti medi sull'intera popolazione. Spesso l'effetto del trattamento varia notevolmente da un individuo o gruppo di individui all'altro. Tali effetti del trattamento sono chiamati eterogenei.

Confondente

Domande importanti sulla causalità riguardano quando possiamo riscrivere un'operazione do in termini di probabilità condizionate. Quando ciò è possibile, possiamo stimare l'effetto dell'operazione do dalle probabilità condizionali convenzionali che possiamo stimare dai dati.

La domanda più semplice di questo tipo chiede quando un effetto causale $P\{Y = y | do(X := x)\}$ coincide con la probabilità della condizione $P\{Y = y | X = x\}$. In generale, questo è

non è vero. Dopotutto, la differenza tra osservazione (probabilità condizionata) e azione (calcolo interventistico) è ciò che ha motivato lo sviluppo della causalità.

Il disaccordo tra affermazioni interventistiche e affermazioni condizionali è così importante che ha un nome ben noto: confondimento. Diciamo che X e Y sono confusi quando l'effetto causale dell'azione $X := x$ su Y non coincide con la corrispondente probabilità condizionata.

Quando X e Y vengono confusi, possiamo chiederci se esiste qualche combinazione di affermazioni di probabilità condizionata che ci danno l'effetto desiderato di un intervento do. Ciò è generalmente possibile dato un grafo causale condizionando sui nodi genitori PA del nodo X:

$$P\{Y = y | do(X := x)\} = \sum_z P\{Y = y | X = x, PA = z\}P\{PA = z\}$$

Questa formula è chiamata formula di aggiustamento. Ci fornisce un modo per stimare l'effetto di un intervento in termini di probabilità condizionate.

La formula di aggiustamento è un esempio di ciò che viene spesso chiamato controllo di un insieme di variabili: stiamo l'effetto di X su Y separatamente in ogni fetta della popolazione definita da una condizione $Z = z$ per ogni possibile valore di z. Quindi calcoliamo la media di questi effetti stimati di sottopopolazione ponderati in base alla probabilità di $Z = z$ nella popolazione. Per fare un esempio, quando controlliamo per età, intendiamo che stiamo un effetto separatamente in ogni possibile fascia di età e poi mediamo i risultati in modo che ogni fascia di età sia ponderata per la frazione della popolazione che rientra in quella fascia di età.

Controllare più variabili in uno studio non è sempre la scelta giusta. Dipende dalla struttura del grafico. Consideriamo cosa succede quando controlliamo la variabile Z nei tre grafici causali di cui abbiamo discusso sopra.

- Il controllo di una variabile di confusione Z in una forcella $X \rightarrow Z \rightarrow Y$ decon- trovato l'effetto di X su Y.
- Il controllo di un mediatore Z su una catena $X \rightarrow Z \rightarrow Y$ ne eliminerà alcuni l'influenza causale di X su Y.
- Il controllo per un collisore creerà una correlazione tra X e Y. Questo è l'opposto di ciò che il controllo per Z realizza nel caso di un fork. Lo stesso vale se controlliamo un discendente di un collisore.

Il criterio della backdoor

A questo punto, potremmo temere che le cose diventino sempre più complicate. Man mano che introduciamo più nodi nel nostro grafico, potremmo temere un'esplosione combinatoria di possibili scenari da discutere. Fortunatamente, esistono criteri semplici e sufficienti per scegliere un insieme di variabili deconfondenti che siano sicure da controllare.

Una nozione ben nota della teoria dei grafi è il criterio della backdoor.¹⁸⁰ Due variabili vengono confuse se esiste un cosiddetto percorso backdoor tra di loro. Un percorso backdoor da X a Y è qualsiasi percorso che inizia da X con un bordo all'indietro "y" in X come:

$$X \rightarrow A \rightarrow B \rightarrow C \rightarrow Y$$

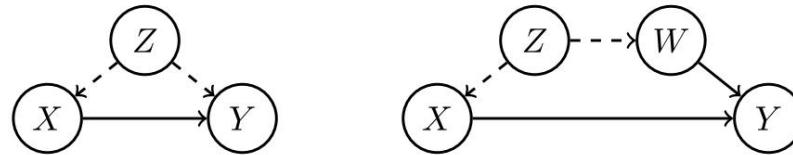


Figura 5.6: Due casi di confondimento non osservato.

Intuitivamente, i percorsi backdoor consentono il flusso di informazioni da X a Y in modo non causale. Per deconfondere una coppia di variabili dobbiamo selezionare un insieme di variabili backdoor che "blocca" tutti i percorsi backdoor tra i due nodi. Un percorso backdoor che coinvolge una catena A → B → C può essere bloccato controllando B. Le informazioni per impostazione predefinita non possono fluire attraverso un collisore A → B → C. Quindi dobbiamo solo stare attenti a non aprire il flusso di informazioni attraverso un collisore condizionandolo sul collisore o discendente di un collisore.

Confondimento non osservato

La formula di aggiustamento potrebbe suggerire che possiamo sempre eliminare i bias confondenti condizionando i nodi principali. Tuttavia, questo è vero solo in assenza di fattori di confusione non osservati. Nella pratica spesso ci sono variabili difficili da misurare o semplicemente non registrate. Possiamo ancora includere tali nodi non osservati in un grafico, in genere denotando la loro influenza con linee tratteggiate, invece che con linee continue.

La figura sopra mostra due casi di confondimento non osservato. Nel primo esempio, l'effetto causale di X su Y non è identificabile. Nel secondo caso, possiamo bloccare il percorso confondente della backdoor X → Z → W → Y controllando per W anche se Z non viene osservato. Il criterio backdoor ci consente di aggirare i fattori confondenti non osservati in alcuni casi in cui la formula di aggiustamento da sola non sarebbe sufficiente.

Il confondimento non osservato rimane tuttavia un grosso ostacolo nella pratica. Il problema non è solo la mancanza di misurazione, ma spesso la mancanza di anticipazione o consapevolezza di una variabile controfondante. Possiamo provare a combattere il confondimento non osservato aumentando il numero di variabili prese in considerazione. Ma quando introduciamo più variabili nel nostro studio, aumentiamo anche l'onere di trovare un modello causale valido per tutte le variabili prese in considerazione. In pratica, non è raro controllare il maggior numero possibile di variabili nella speranza di disattivare i bias confondenti. Tuttavia, come abbiamo visto, il controllo dei mediatori o dei collisori può essere dannoso.

Randomizzazione

Il criterio backdoor fornisce un modo non sperimentale per eliminare i bias confondenti, dato un modello causale e una quantità sufficiente di dati osservativi dalla distribuzione congiunta delle variabili. Un metodo sperimentale alternativo per eliminare i bias confondenti è il noto studio randomizzato e controllato.

In uno studio randomizzato e controllato un gruppo di soggetti viene suddiviso casualmente in un gruppo di controllo e un gruppo di trattamento. I partecipanti non sanno a quale gruppo appartengono

sono stati assegnati e nemmeno il personale che gestisce lo studio. Il gruppo di trattamento riceve un trattamento vero e proprio, come un farmaco di cui viene testata l'efficacia, mentre il gruppo di controllo riceve un placebo identico nell'aspetto. Una variabile di risultato viene misurata per tutti i soggetti.

L'obiettivo di uno studio randomizzato e controllato è quello di rompere l'inclinazione naturale. Invece di osservare chi ha scelto di farsi curare da solo, assegniamo il trattamento in modo casuale. Pensando in termini di modelli causali, ciò significa che eliminiamo tutti i margini in entrata nella variabile di trattamento. In particolare, questo chiude tutti i percorsi backdoor e quindi evita pregiudizi confondenti.

Ci sono molte ragioni per cui spesso gli studi randomizzati e controllati sono difficili o impossibili da gestire. Il trattamento potrebbe essere fisicamente o legalmente impossibile, troppo costoso o troppo pericoloso. Come abbiamo visto, gli studi randomizzati e controllati non sono sempre necessari per evitare bias confondenti e per ragionare su causa ed effetto.

Né sono esenti da problemi e insidie.¹⁸¹

Analisi grafica della discriminazione

Esploriamo ora come possiamo utilizzare i grafici causali nelle discussioni sulla discriminazione . Ritroneremo all'esempio delle ammissioni dei laureati a Berkeley e svilupperemo una prospettiva causale sull'analisi precedente.

Il primo passo è elaborare un grafico causale plausibile coerente con i dati che abbiamo visto in precedenza. I dati contenevano solo tre variabili, sesso A, scelta del dipartimento Z e decisione di ammissione Y. Ha senso tracciare due frecce A → Y e Z → Y, perché entrambe le caratteristiche A e Z sono disponibili per l'istituto al momento della decisione di ammissione . Per ora disegneremo un'altra freccia, semplicemente perché è necessario. Se includessimo solo le due frecce A → Y e Z → Y, il nostro grafico affermerebbe che A e Z sono statisticamente indipendenti. Tuttavia, questa affermazione non è coerente con i dati. Dalla tabella possiamo vedere che diversi dipartimenti presentano un pregiudizio di genere statisticamente significativo tra i candidati. Ciò significa che dobbiamo includere la freccia A → Z o Z → A. Decidere tra le due non è così semplice come potrebbe sembrare a prima vista.

Se interpretassimo A nel senso più stretto possibile come il sesso dichiarato dal richiedente, ovvero letteralmente quale casella ha selezionato nel modulo di domanda, potremmo immaginare uno scenario in cui alcuni richiedenti scelgono di (erroneamente) dichiarare il proprio sesso in un certo modo che dipende in parte dalla scelta del dipartimento. Anche supponendo che non si verifichino segnalazioni errate, è difficile dimostrare che il sesso segnalato sia una causa plausibile della scelta del dipartimento. Il fatto che un candidato abbia spuntato una casella etichettata come maschio non è certamente la causa del suo interesse per l'ingegneria.

La storia causale proposta nello studio è diversa. Allude a un processo di socializzazione e formazione delle preferenze che ha avuto luogo nella vita del richiedente prima che facesse domanda. È questo processo che, almeno in parte, dipendeva dal sesso del richiedente. Per allineare questa storia con il nostro grafico causale, abbiamo bisogno che la variabile A faccia riferimento a qualunque entità ontologica sia che attraverso questo "processo di socializzazione" influenza le preferenze intellettuali e professionali e, quindi, la scelta del dipartimento.

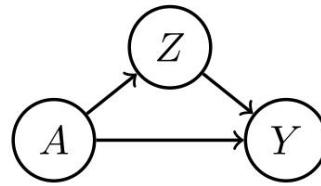


Figura 5.7: Possibile grafico causale per lo scenario di ammissione dei laureati alla UC Berkeley.

È difficile sostenere che questa entità ontologica coincida con il sesso come tratto biologico. Non esiste alcuna base scientifica per sostenere che il tratto biologico sessuale sia ciò che determina le nostre preferenze intellettuali. Pochi studiosi (se ce ne sono) attualmente tenterebbero di sostenere un'affermazione secondo cui due cromosomi X suscitano interesse per la letteratura inglese.

La verità è che non conosciamo l'esatto meccanismo con cui la cosa a cui fa riferimento A influenza la scelta del dipartimento. Disegnando la freccia dalla A alla Z affermiamo – forse con una certa ingenuità o ignoranza – che esiste un tale meccanismo. Discuteremo più avanti in modo approfondito l'importante difficoltà che abbiamo incontrato qui. Per ora ci impegniamo in questa scelta modellistica e arriviamo così al grafico seguente.

In questo grafico, la scelta del dipartimento media l'influenza del genere sulle ammissioni. C'è un percorso diretto da A a Y e un percorso indiretto che passa per Z.

Utilizzeremo questo modello per fare pressione sull'affermazione secondo cui non esiste alcuna prova di discriminazione sessuale. Nel linguaggio causale, l'argomentazione aveva due componenti:

1. Non sembra esserci alcun effetto diretto del sesso A sulla decisione di ammissione Y che favorisce gli uomini.
2. L'effetto indiretto di A su Y mediato dalla scelta del dipartimento dovrebbe non possono essere considerati prova di discriminazione.

Discuteremo entrambi gli argomenti uno dopo l'altro.

Effetti diretti

Per ottenere l'effetto diretto di A su Y dobbiamo disabilitare tutti i percorsi tra A e Y ad eccezione del collegamento diretto. Nel nostro modello, possiamo ottenere questo risultato mantenendo costante la scelta del dipartimento Z e valutando la distribuzione condizionale di Y dato A. Ricordiamo che mantenere costante una variabile generalmente non equivale a condizionare la variabile. Nello specifico, sorgerebbe un problema se la scelta del dipartimento e l'esito dell'ammissione fossero confusi da un'altra variabile, come lo stato di residenza R

La scelta del dipartimento è ora un collisore tra A e R. Il condizionamento su un collisore apre il percorso backdoor A → Z → R → Y. In questo grafico, il condizionamento sulla scelta del dipartimento non ci dà l'effetto diretto desiderato. La reale possibilità che lo stato di residenza confonda la scelta e la decisione del dipartimento è stata oggetto di uno scambio tra Bickel e Kruskal.¹⁸²

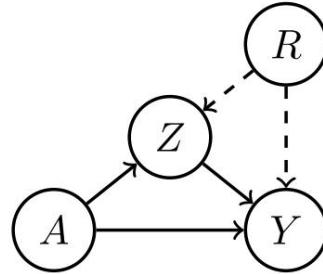


Figura 5.8: Grafico causale alternativo per lo scenario di ammissione dei laureati dell'UC Berkeley che mostra l'influenza della residenza.

Se assumiamo, tuttavia, che la scelta del dipartimento e le decisioni di ammissione non siano confuse, allora l'approccio adottato da Bickel, Hammel e O'Connell supporta effettivamente la prima affermazione. Sfortunatamente, l'effetto diretto di una variabile protetta su una decisione è di per sé una misura inadeguata della discriminazione. A livello tecnico è piuttosto fragile in quanto non è in grado di rilevare alcuna forma di discriminazione per procura. Il dipartimento potrebbe, ad esempio, utilizzare la dichiarazione personale del richiedente per fare deduzioni sul suo genere, che verranno poi utilizzate per discriminare.

Possiamo pensare all'effetto diretto come corrispondente all'uso esplicito dell'attributo nella regola decisionale. L'assenza di un effetto diretto corrisponde vagamente alla nozione alquanto travagliata di una regola decisionale cieca che non ha accesso esplicito all'attributo sensibile. Come abbiamo sostenuto nei capitoli precedenti, le regole della decisione cieca possono ancora essere la base di pratiche discriminatorie.

Percorsi indiretti

Passiamo all'effetto indiretto del sesso sull'ammissione che passa attraverso la scelta del dipartimento. È forte la tentazione di pensare al nodo Z come un riferimento alle preferenze intrinseche del dipartimento del richiedente. In quest'ottica, il dipartimento non è responsabile delle preferenze del richiedente. Pertanto l'influenza mediatrice delle preferenze dipartimentali non viene interpretata come un segno di discriminazione. Questo, tuttavia, è un giudizio sostanziale che potrebbe non corrispondere a un dato di fatto. Esistono altri scenari plausibili coerenti sia con i dati che con il nostro modello causale, in cui il percorso indiretto codifica un modello di discriminazione.

Ad esempio, il comitato di ammissione potrebbe aver pubblicizzato il programma in un modo tale da scoraggiare fortemente le donne dal candidarsi. In questo caso, la preferenza del dipartimento misura in parte l'esposizione a questa campagna pubblicitaria ostile. In alternativa, il dipartimento potrebbe avere precedenti di comportamenti ostili contro le donne ed è la consapevolezza di ciò che modella le preferenze di un richiedente. Infine, anche pratiche palesemente discriminatorie, come quella di retribuire le donne a un tasso inferiore rispetto agli studenti laureati di pari qualifica, corrispondono a un effetto indiretto mediato dalla scelta del dipartimento.

Accettare il percorso indiretto come non discriminatorio significa affermare che tutti gli scenari che abbiamo descritto sono ritenuti non plausibili. Fondamentalmente, ci confrontiamo

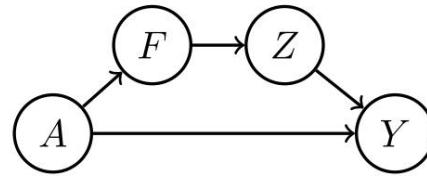


Figura 5.9: Grafico causale alternativo per lo scenario di ammissione dei laureati dell'UC Berkeley in cui le preferenze del dipartimento sono modellate dalla paura della discriminazione.

una questione sostanziale. Il percorso A → Z → Y potrebbe essere il luogo in cui si verifica la discriminazione o ciò che ne spiega l'assenza. Il caso in cui ci troviamo non è una questione puramente tecnica e non può essere risolto senza la conoscenza dell'argomento. La modellazione causale ci fornisce un quadro per esporre queste domande, ma non necessariamente per risolverle.

Ispezione del percorso

Per riassumere, la discriminazione può non verificarsi solo nel percorso diretto dalla categoria sensibile al risultato. Percorsi di mediazione apparentemente innocui possono nascondere pratiche discriminatorie. Dobbiamo discutere attentamente quali percorsi consideriamo prove a favore o contro la discriminazione.

Per apprezzare questo punto, confrontiamo il nostro scenario di Berkeley con l'importante caso legale Griggs contro Duke Power Co. che fu discusso davanti alla Corte Suprema degli Stati Uniti nel 1970. La Duke Power Company aveva introdotto il requisito di un diploma di scuola superiore per alcuni lavori più retribuiti. Potremmo tracciare un grafico causale per questo scenario non dissimile da quello del caso Berkeley. C'è una variabile di mediazione (qui, livello di istruzione), una categoria sensibile (qui, razza) e un risultato occupazionale (qui, impiego in un lavoro più retribuito). L'azienda non prendeva direttamente decisioni sull'assunzione in base alla razza, ma utilizzava piuttosto la variabile di mediazione.

La corte ha stabilito che il requisito di un diploma di scuola superiore non era giustificato da necessità aziendali, ma piuttosto aveva un impatto negativo sui gruppi etnici minoritari dove la prevalenza dei diplomi di scuola superiore è inferiore. In altre parole, la corte ha deciso che l'uso di questa variabile mediatrice non era un argomento contro, ma piuttosto a favore della discriminazione.

Glymour¹⁸³ sottolinea un altro punto correlato e importante sulle caratteristiche morali ter dell'analisi della mediazione:

Implicitamente, la questione di cosa media gli effetti sociali osservati informa la nostra visione di quali tipi di disuguaglianze sono socialmente accettabili e quali tipi richiedono una correzione da parte delle politiche sociali. Ad esempio, la conclusione che le donne sono "biologicamente programmate" per essere depresse più degli uomini può alleviare l'obbligo sociale di cercare di ridurre le disuguaglianze di genere nella depressione. Tuttavia, se le persone si deprimono ogni volta che, ad esempio, vengono molestate sessualmente – e le donne sono più frequentemente molestate sessualmente rispetto agli uomini – ciò suggerisce un obbligo sociale molto forte verso

ridurre la disparità nella depressione riducendo la disparità nelle molestie sessuali.

Concludendo con una nota tecnica, al momento non disponiamo di un metodo per stimare gli effetti indiretti. La stima di un effetto indiretto richiede in qualche modo di disabilitare l'influenza diretta. Non c'è modo di farlo con l'operazione do che abbiamo visto finora. Tuttavia, introdurremo a breve i controfattuali, che, tra le altre applicazioni, ci forniranno un modo per stimare gli effetti specifici del percorso.

Discriminazione strutturale

C'è un ulteriore problema che finora abbiamo trascurato. Immaginate un'amministrazione universitaria dispettosa che tagli sistematicamente i fondi ai programmi di laurea che attirano più candidate donne. Questo modello strutturale di discriminazione è invisibile dal modello causale che abbiamo disegnato. C'è una sorta di mancata corrispondenza del tipo qui. Il nostro modello parla dei singoli candidati, delle loro preferenze di dipartimento e dei loro risultati. In altre parole, gli individui sono le unità della nostra indagine. La politica universitaria non è uno dei meccanismi che il nostro modello mette in luce. Non possiamo caratterizzare la politica universitaria per farne un attributo dell'individuo. Di conseguenza, nel nostro modello non possiamo parlare di politica universitaria come causa di discriminazione.

Il modello che abbiamo scelto ci impegna in una prospettiva individualistica che inquadra la discriminazione come conseguenza del modo in cui i decisori rispondono alle informazioni sugli individui. È utile un'analogia. In epidemiologia, gli scienziati possono ricercare la causa dei risultati sanitari negli aspetti biomedici e nelle scelte di vita degli individui, ad esempio se un individuo fuma o meno, fa esercizio fisico, mantiene una dieta equilibrata, ecc. Il crescente campo dell'epidemiologia sociale critica la visione delle scelte individuali come principali cause dei risultati sanitari, e richiama invece l'attenzione sulle cause sociali e strutturali,¹⁸⁴ come la povertà e la disuguaglianza.

Allo stesso modo, possiamo contrapporre la prospettiva individualistica sulla discriminazione alla discriminazione strutturale. In linea di principio, la modellazione causale può essere utilizzata anche per studiare le cause della discriminazione strutturale. Ma richiede una prospettiva diversa da quella che abbiamo scelto per il nostro scenario di Berkeley.

Controfattuali

I modelli causal strutturali completamente specificati ci consentono di porre domande causali che sono più delicate del semplice effetto di un'azione. Nello specifico, possiamo porre domande controfattuali come: avrei evitato l'ingorgo se stamattina avessi preso una strada diversa? Le domande controfattuali sono comuni e rilevanti per le questioni di discriminazione. Possiamo computerizzare la risposta a domande controfattuali dato un modello causale strutturale. La procedura per estrarre la risposta dal modello sembra inizialmente un po' complessa. Esamineremo i dettagli formali partendo da un semplice esempio prima di tornare alla nostra discussione sulla discriminazione.

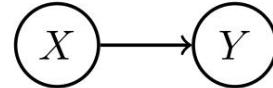


Figura 5.10: Diagramma causale per il nostro scenario di traffico.

Un semplice controfattuale

Per comprendere i controfattuali, dobbiamo prima convincerci che non sono così semplici come una singola sostituzione nel nostro modello.

Supponiamo che ogni mattina dobbiamo decidere tra due percorsi $X = 0$ e $X = 1$. Nei giorni di cattivo traffico, indicati da $U = 1$, entrambi i percorsi sono cattivi. Nelle giornate buone, indicate da $U = 0$, il traffico su entrambi i percorsi è buono a meno che non si sia verificato un incidente sul percorso. Diciamo che $U \sim B(1/2)$ segue la distribuzione di un lancio imparziale di una moneta. Gli incidenti si verificano indipendentemente su entrambi i percorsi con probabilità $1/2$.

Quindi, scegliamo due variabili casuali di Bernoulli $U_0, U_1 \sim B(1/2)$ che ci dicono se c'è un incidente rispettivamente sulla rotta 0 e sulla rotta 1. Rifiutiamo ogni orientamento esterno e decidiamo invece quale percorso prendere in modo uniforme e casuale. Cioè, anche $X := UX \sim B(1/2)$ è un lancio imparziale della moneta.

Introduciamo una variabile $Y \in \{0, 1\}$ che ci dica se il traffico sul percorso scelto è buono ($Y = 0$) o cattivo ($Y = 1$). Riflettendo la nostra discussione sopra, possiamo esprimere Y come

$$Y := X \cdot \max\{U, U_1\} + (1 - X) \max\{U, U_0\}.$$

In parole povere quando $X = 0$ il primo termine scompare e quindi il traffico è determinato dal maggiore dei due valori U e U_0 . Allo stesso modo, quando $X = 1$ il traffico è determinato dal maggiore tra U e U_1 .

Ora, supponiamo che una mattina abbiamo $X = 1$ e osserviamo cattivo traffico $Y = 1$.

Sarebbe stato meglio prendere la strada alternativa stamattina?

Un tentativo naturale di rispondere a questa domanda è calcolare la probabilità che $Y = 0$ dopo l'operazione $X := 0$, cioè $PM[X:=0] (Y = 0)$. Un rapido calcolo rivela che questa probabilità è $= 1/4$. Infatti, data la sostituzione $X := \frac{U_1}{U_0+U_1}$ nel nostro modello, affinché il traffico sia buono abbiamo bisogno che $\max\{U, U_0\} = 0$. Ciò può accadere solo quando sia $U = 0$ (probabilità $1/2$) che $U_0 = 0$ (probabilità $1/2$).

Ma questa non è la risposta corretta alla nostra domanda. Il motivo è che abbiamo preso il percorso $X = 1$ e osservato che $Y = 1$. Da questa osservazione, possiamo dedurre che alcune condizioni di fondo non si sono manifestate perché non sono coerenti con il risultato osservato. Formalmente ciò significa che determinate impostazioni delle variabili di rumore (U, U_0, U_1) non sono più fattibili dato l'evento osservato $\{Y = 1, X = 1\}$.

Nello specifico, se U e U_1 fossero stati entrambi pari a zero, non avremmo visto alcun traffico negativo sul percorso $X = 1$, ma questo è contrario alla nostra osservazione. Infatti, l'evidenza disponibile $\{Y = 1, X = 1\}$ lascia solo le seguenti impostazioni per U e U_1 :

Tabella 5.2: Possibili impostazioni del rumore dopo aver osservato le prove

U	U1
0	1
1	1
10	

Tralasciamo U_0 dalla tabella poiché la sua distribuzione non è influenzata dalla nostra osservazione. Ciascuno dei restanti tre casi è ugualmente probabile, il che significa in particolare che l'evento $U = 1$ ora ha probabilità $2/3$. In assenza di ulteriori prove, ricordiamo che $U = 1$ aveva probabilità $1/2$. Ciò significa che l'evidenza osservata $\{Y = 1, X = 1\}$ ha distorto la distribuzione della variabile rumore U verso 1. Usiamo la lettera U per riferirci a questa versione distorta di U .

Formalmente U è distribuito secondo la distribuzione di U condizionata all'evento $\{Y = 1, X = 1\}$.

Lavorando con questa variabile di rumore distorta, possiamo nuovamente considerare l'effetto dell'azione $X := 0$ sul risultato Y . Per $Y = 0$ abbiamo bisogno che $\max\{U, U_0\} = 0$. Ciò significa che $U = 0$, un evento che ora ha probabilità $1/3$, e $U_0 = 0$ (probabilità $1/2$ come prima). Quindi, otteniamo la probabilità $1/6 = 1/2 \cdot 1/3$ per l'evento che $Y = 0$ con la nostra operazione do $X := 0$, e dopo aver aggiornato le variabili di rumore per tenere conto dell'osservazione $\{Y = 1, X = 1\}$.

Per riassumere, incorporando le prove disponibili nel nostro calcolo è diminuita la probabilità di assenza di traffico ($Y = 0$) quando si sceglie il percorso 0 da $1/4$ a $1/6$. La ragione intuitiva è che le prove rendevano più probabile che in generale si trattasse di una giornata con traffico intenso e che anche il percorso alternativo sarebbe stato intasato. Più formalmente, l'evento che abbiamo osservato distorce la distribuzione delle variabili di rumore esogeno.

Pensiamo al risultato che abbiamo appena calcolato come al controfattuale della scelta del percorso alternativo dato che il percorso che abbiamo scelto presentava scarso traffico.

La ricetta generale

Possiamo generalizzare la nostra discussione sul calcolo dei controfattuali dall'esempio precedente a una procedura generale. I passaggi essenziali erano tre. Innanzitutto, abbiamo incorporato le prove osservative disponibili influenzando le variabili di rumore esogeno attraverso un'operazione di condizionamento. In secondo luogo, abbiamo eseguito un'operazione do nel modello causale strutturale dopo aver sostituito le variabili di rumore distorte. In terzo luogo, abbiamo calcolato la distribuzione di una variabile target. Questi tre passaggi sono tipicamente chiamati rapimento, azione e previsione, come può essere descritto come segue.

Definizione 5. Dato un modello causale strutturale M , un evento osservato E , un'azione $X := x$ e la variabile obiettivo Y , definiamo il controfattuale $Y_{X:=x}(E)$ mediante la seguente procedura in tre passaggi:

1. **Abduzione:** regolare le variabili del rumore in modo che siano coerenti con l'evento osservato.

Formalmente, condizionare la distribuzione congiunta di $U = (U_1, \dots, U_d)$ all'evento E . Ciò si traduce in una distribuzione distorta U .

2. **Azione:** eseguire l'intervento $X := x$ nel modello causale strutturale M risultante
nel modello $M = M[X := x]$.
3. **Previsione:** calcolare il controfattuale target $YX:=x(E)$ utilizzando U come casuale
seme in M .

È importante rendersi conto che questa procedura definisce cosa sia un controfattuale in un modello causale strutturale. La notazione $YX:=x(E)$ denota l'esito della procedura e fa parte della definizione. Non abbiamo mai incontrato questa notazione prima. In parole poche, interpretiamo il controfattuale formale $YX:=x(E)$ come il valore che Y avrebbe assunto se la variabile X fosse stata impostata sul valore x nelle circostanze descritte dall'evento E .

In generale, il controfattuale $YX:=x(E)$ è una variabile casuale che varia con U . Ma i controfattuali possono anche essere deterministici. Quando l'evento E restringe la distribuzione di U ad un unico punto massa, detto unità, la variabile U è costante e quindi il controfattuale $YX:=x(E)$ si riduce ad un unico numero. In questo caso, è comune usare la notazione abbreviata $Yx(u) = YX:=x(\{U = u\})$, dove rendiamo implicita la variabile X e facciamo riferimento a una singola unità.

La motivazione per il nome unità deriva dalla situazione comune in cui il modello causale strutturale descrive una popolazione di entità che formano le unità atomiche del nostro studio. È comune che un'unità sia un individuo (o la descrizione di un singolo individuo). Tuttavia, a seconda dell'applicazione, la scelta delle unità può variare. Nel nostro esempio di traffico, le variabili del rumore determinano quale percorso prendiamo e quali sono le condizioni stradali.

Le risposte alle domande controfattuali dipendono fortemente dalle specificità del modello causale strutturale, compreso il modello preciso di come entrano in gioco le variabili esogene del rumore. È possibile costruire due modelli che hanno strutture grafiche identiche e si comportano in modo identico in caso di interventi, ma danno risposte diverse a domande controfattuali.¹⁸⁵

Risultati potenziali

Il quadro dei risultati potenziali è una base formale popolare per l'inferenza causale, che tratta i controfattuali in modo diverso. Invece di derivarli da un modello causale strutturale, assumiamo la loro esistenza come variabili casuali ordinarie, anche se alcune non osservate.

Nello specifico, assumiamo che per ogni unità u esistano variabili casuali $Yx(u)$ per ogni possibile valore dell'assegnazione x . Nel modello dei risultati potenziali, è consuetudine pensare a una variabile di trattamento binaria X in modo che x assuma solo due valori, 0 per non trattato e 1 per trattato. Questo ci dà due potenziali variabili di risultato $Y0(u)$ e $Y1(u)$ per ciascuna unità u . C'è qualche potenziale confusione tra le notazioni qui. I lettori che hanno familiarità con il modello dei risultati potenziali possono essere abituati alla notazione " $Yi(0), Yi(1)$ " per i due risultati potenziali corrispondenti all'unità i .

Nella nostra notazione l'unità (o, più in generale, l'insieme di unità) appare tra parentesi e il pedice indica il valore sostituito per la variabile su cui si interviene.

Il punto chiave del modello dei risultati potenziali è che osserviamo solo il risultato potenziale $Y1(u)$ per le unità trattate. Per le unità non trattate noi

osservare $Y_0(u)$. In altre parole, però, non potremo mai osservarli entrambi contemporaneamente si presume che esistano entrambi in senso formale. Formalmente, il risultato $Y(u)$ per l'unità u che osserviamo dipende dal trattamento binario $T(u)$ ed è data da espressione:

$$Y(u) = Y_0(u) \cdot (1 \circ T(u)) + Y_1(u) \cdot T(u)$$

Spesso è conveniente omettere le parentesi dalla nostra notazione per le controfactuals in modo che questa espressione sia $Y = Y_0 \cdot (1 \circ T) + Y_1 \cdot T$.

Possiamo rivisitare il nostro esempio di traffico in questo contesto. La tabella successiva riassume quali informazioni sono osservabili nel modello dei risultati potenziali. Pensiamo al percorso che scegliamo come variabile di trattamento e il traffico osservato come variabile riflettente dei due possibili risultati.

Tabella 5.3: Esempio di traffico nel modello dei risultati potenziali

Percorso X	Risultato Y0	Risultato Y1	Probabilità
0	0	?	1/8
0	1	?	3/8
1	?	0	1/8
1	?	1	3/8

Spesso queste informazioni si presentano sotto forma di campioni. Ad esempio, potremmo osservare il traffico in giorni diversi. Con un numero sufficiente di campioni, possiamo stimare le frequenze di cui sopra con precisione arbitraria.

Tabella 5.4: Dati sul traffico nel modello dei risultati potenziali

Giorno	Percorso X	Risultato Y0	Risultato Y1
1	0	1	?
2	0	0	?
3	1	?	1
4	0	1	?
5	1	?	0
...

Una query tipica nel modello dei risultati potenziali è l'effetto medio del trattamento $E[Y_1 \circ Y_0]$. Qui l'aspettativa viene rilevata nelle unità adeguatamente ponderate nel nostro studio. Se le unità corrispondono a individui con lo stesso peso, l'aspettativa è una media rispetto a questi individui.

Nel nostro esempio di traffico originale, c'erano 16 unità corrispondenti alle condizioni di fondo date dalle quattro variabili binarie U, U_0, U_1, UX . Quando le unità nel modello dei risultati potenziali concordano quindi con quelli di un modello causale strutturale gli effetti causali calcolati nel modello dei risultati potenziali concordano con quelli com-

inserito nel modello delle equazioni strutturali. I due quadri formali sono perfettamente coerenti tra loro.

Come è intuitivo dalla tabella sopra, l'inferenza causale nel quadro dei risultati potenziali può essere pensata come se riempisse le voci mancanti ("?") nella tabella sopra. Questa operazione viene talvolta chiamata imputazione dei dati mancanti e esistono numerosi metodi statistici per questo compito. Se potessimo rivelare cosa si nasconde dietro i punti interrogativi, stimare l'effetto medio del trattamento sarebbe facile come contare le righe.

Esiste una serie di condizioni stabilitate sotto le quali l'inferenza causale diventa possibile:

1. **Presupposizione del valore del trattamento dell'unità stabile (SUTVA):** il trattamento ricevuto da un'unità non modifica l'effetto del trattamento per qualsiasi altra unità.
2. **Consistenza:** Formalmente, $Y = Y_0(1 \circ T) + Y_1T$. Cioè, $Y = Y_0$ se $T = 0$ e $Y = Y_1$ se $T = 1$. In parole, il risultato Y concorda con il risultato potenziale corrispondente all'indicatore di trattamento.
3. **Ignorabilità:** i risultati potenziali sono indipendenti dal trattamento date alcune variabili deconfondenti Z , cioè $T \circ (Y_0, Y_1) | Z$. In parole povere, i risultati potenziali sono condizionatamente indipendenti dal trattamento, dato un insieme di variabili decongestionanti.

Le prime due assunzioni valgono automaticamente per variabili controllattuali derivate da modelli causali strutturali secondo la procedura sopra descritta. Ciò presuppone che le unità nel quadro dei risultati potenziali corrispondano ai valori atomici delle variabili di fondo nel modello causale strutturale.

La terza ipotesi è importante. È più semplice pensarla come se mirasse a formalizzare le garanzie di uno studio randomizzato e controllato perfettamente eseguito. L'ipotesi di per sé non può essere verificata o falsificata, dal momento che non abbiamo mai accesso a campioni con entrambi i potenziali risultati manifestati. Tuttavia, possiamo verificare se l'ipotesi è coerente con un dato modello causale strutturale controllando se l'insieme Z blocca tutti i percorsi backdoor dal trattamento T al risultato Y .

Non c'è tensione tra modelli causali strutturali e risultati potenziali e non c'è nulla di male nell'avere familiarità con entrambi. È comunque opportuno spendere qualche parola sulle differenze tra i due approcci.

Possiamo derivare risultati potenziali da un modello causale strutturale come abbiamo fatto sopra, ma non possiamo derivare un modello causale strutturale solo dai risultati potenziali. Un modello causale strutturale in generale codifica più ipotesi sulle relazioni tra le variabili. Ciò ha diverse conseguenze. Da un lato, un modello causale strutturale ci fornisce un insieme più ampio di concetti formali (grafici causali, percorsi di mediazione, controllattuali per ogni variabile e così via). D'altra parte, elaborare un modello causale strutturale plausibilmente valido è spesso un compito arduo che potrebbe richiedere conoscenze che semplicemente non sono disponibili. Di seguito approfondiremo le questioni relative alla validità. La difficoltà di elaborare un modello causale plausibile spesso espone questioni sostanziali irrisolte che richiedono prima una risoluzione.

Il modello dei risultati potenziali, al contrario, è generalmente più facile da applicare. Esiste un'ampia gamma di stimatori statistici degli effetti causali che possono essere facilmente applicati ai dati osservativi. Ma la facilità di applicazione può portare anche ad abusi. IL

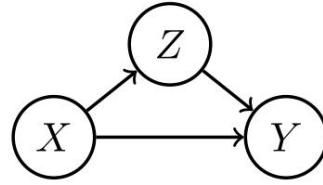


Figura 5.11: Grafico causale con il mediatore Z.

le ipotesi alla base della validità di tali stimatori non sono verificabili sperimentalmente. L'applicazione frivola degli stimatori dell'effetto causale in situazioni in cui i presupposti cruciali non sono validi può portare a risultati falsi e di conseguenza a interventi inefficaci o dannosi.

Analisi controllattuale della discriminazione

I controllattuali per noi servono almeno a due scopi. Dal punto di vista tecnico, i controllattuali ci danno un modo per calcolare gli effetti causali specifici del percorso. Ciò ci consente di rendere l'analisi del percorso una questione quantitativa. Dal punto di vista concettuale, i controllattuali ci consentono di affrontare l'importante dibattito normativo sulla questione se la discriminazione possa essere catturata da criteri controllattuali. Discuteremo ciascuno di questi a turno.

Analisi quantitativa del percorso

L'analisi della mediazione è un argomento venerabile che risale a decenni fa.¹⁸⁶ In generale, l'obiettivo dell'analisi della mediazione è identificare un meccanismo attraverso il quale una causa ha un effetto. Esamineremo alcuni sviluppi recenti e il modo in cui si collegano alle questioni di discriminazione.

Nel linguaggio del nostro quadro formale, l'analisi della mediazione mira a scomporre un effetto causale totale in componenti specifiche del percorso. Illustreremo i concetti nel caso base delle tre variabili di un mediatore, sebbene le idee si estendano a strutture più complicate.

Esistono due percorsi diversi da X a Y. Un percorso diretto e un percorso attraverso il mediatore Z. L'aspettativa condizionale $E[Y | X = x]$ raggruppa l'influenza di entrambi i percorsi. Se ci fosse un'altra variabile di confondimento nel nostro grafico che influenza sia X che Y, allora l'aspettativa condizionale includerebbe anche qualunque correlazione sia il risultato del confondimento. Possiamo eliminare il percorso di confusione in virtù dell'operatore do $E[Y | \text{fare}(X := x)]$. Questo ci dà l'effetto totale dell'azione $X := x$ su Y. Ma l'effetto totale fonde ancora i due percorsi causali, l'effetto diretto e l'effetto indiretto. Vedremo ora come individuare separatamente gli effetti diretti e quelli indiretti.

Dell'effetto diretto abbiamo già parlato in precedenza in quanto non richiede alcun controllattuale. Ricordiamo che possiamo mantenere il mediatore fisso al livello $Z := z$ e considerare l'effetto del trattamento $X := 1$ rispetto a nessun trattamento $X := 0$ come segue:

$$\mathbf{E} [Y | \text{do}(X := 1, Z := z)] \circ \mathbf{E} [Y | \text{fare}(X := 0, Z := z)] .$$

Possiamo riscrivere questa espressione in termini di controfattuali in modo equivalente come:

$$\mathbf{E} [Y_{X:=1, Z:=z} \circ Y_{X:=0, Z:=z}] .$$

Per essere chiari, nei nostri modelli causali strutturali l'aspettativa viene presa in considerazione rispetto alle variabili di fondo. In altre parole, i controfattuali all'interno dell'aspettativa vengono invocati con un'impostazione elementare u delle variabili di background, ovvero $Y_{X:=1, Z:=z}(u) \circ Y_{X:=0, Z:=x}(u)$ e la media delle aspettative su tutte le possibili impostazioni.

La formula per l'effetto diretto di cui sopra è solitamente chiamata effetto diretto controllato, poiché richiede l'impostazione della variabile mediatrice a un livello specificato. A volte è desiderabile consentire alla variabile mediatrice di variare come sarebbe avvenuto se non fosse stato effettuato alcun trattamento. Anche questo è possibile con i controfattuali e porta a una nozione chiamata effetto diretto naturale, definito come:

$$\mathbf{E} [Y_{X:=1, Z:=ZX:=0} \circ Y_{X:=0, Z:=ZX:=0}] .$$

Il controfattuale $Y_{X:=1, Z:=ZX:=0}$ è il valore che Y otterrebbe se X fosse stato impostato su 1 e se Z fosse stato impostato sul valore che Z avrebbe assunto se X fosse stato impostato su 0.

Il vantaggio di questa costruzione un po' stravagante è che ci dà un nozione analoga di effetto indiretto naturale:

$$\mathbf{E} [Y_{X:=0, Z:=ZX:=1} \circ Y_{X:=0, Z:=ZX:=0}] .$$

Qui manteniamo costante la variabile di trattamento al livello $X := 0$, ma lasciamo che la variabile mediatrice cambi al valore che avrebbe raggiunto se si fosse verificato il trattamento $X := 1$.

Nel nostro esempio di tre nodi, l'effetto di X su Y non è confuso. In assenza di confusione, l'effetto indiretto naturale corrisponde alla seguente affermazione di probabilità condizionata (che non coinvolge né controfattuali né interventi):

$$\circ z \mathbf{E} [Y | X = 0, Z = z] P(Z = z | X = 1) \circ P(Z = z | X = 0) .$$

In questo caso, possiamo stimare l'effetto naturale diretto e indiretto dai dati osservativi.

Le possibilità tecniche vanno oltre il caso qui discusso. In linea di principio, i controfattuali ci consentono di calcolare tutti i tipi di effetti specifici del percorso anche in presenza di fattori confondenti (osservati). Possiamo anche progettare regole decisionali che eliminino gli effetti specifici del percorso che ritieniamo indesiderabili.

Criteri di discriminazione controfattuale

Oltre alla loro applicazione all'analisi del percorso, i controfattuali possono anche essere utilizzati come strumento per proporre criteri di equità normativa. Considera la configurazione tipica di

Capitolo 3. Abbiamo caratteristiche X, un attributo sensibile A, una variabile di risultato Y e un predittore Y.

Un criterio tecnicamente naturale direbbe quanto segue: per ogni possibile demografia descritta dall'evento $E := \{X := x, A := a\}$ e ogni possibile impostazione a di A chiediamo che il controfattuale $Y_A:=a(E)$ e il controfattuale $Y_{A:=a}(E)$ seguono la stessa distribuzione.

Introdotta come equità controfattuale,¹⁸⁷ ci riferiamo a questa condizione come parità demografica controfattuale, poiché è strettamente correlata al criterio osservativo di parità demografica condizionale. Ricordiamo che la parità demografica condizionale richiede che in ciascun gruppo demografico definito da un'impostazione di funzionalità $X = x$, l'attributo sensibile sia indipendente dal predittore. Formalmente abbiamo la relazione di indipendenza condizionale $Y \perp\!\!\!\perp A | X$. Nel caso di un predittore binario, questa condizione equivale a richiedere per tutte le impostazioni delle funzionalità x e i gruppi a, a :

$$E[Y | X = x, A = a] = E[Y | X = x, A = a]$$

Il modo più semplice per soddisfare la parità demografica controfattuale è che il predittore Y utilizzi solo i non discendenti di A nel grafico causale. Ciò è analogo alla condizione statistica di utilizzare solo caratteristiche indipendenti da A.

Nello stesso modo in cui abbiamo definito un analogo controfattuale della parità demografica, possiamo esplorare analogie causali di altri criteri statistici nel Capitolo 3. Nel fare ciò, dobbiamo stare attenti nel separare le questioni tecniche sulla differenza tra criteri osservativi e causali da quelli contenuto normativo del criterio. Solo perché una variante causale di un criterio potrebbe aggirare alcuni problemi statistici di correlazioni non causali non significa che il criterio causale risolva preoccupazioni o questioni normative con il suo cugino osservativo.

I controfattuali nel diritto

Gratteremo ora la superficie di un argomento profondo della dottrina giuridica sul quale ritorneremo nel Capitolo 6 dopo aver sviluppato una maggiore familiarità con il contesto giuridico. L'argomento è il rapporto tra pretese controfattuali causali e casi giuridici di discriminazione. Molti studiosi tecnici vedono il sostegno a un'interpretazione controfattuale della legge sulla discriminazione degli Stati Uniti in varie sentenze di giudici che sembravano aver invocato un linguaggio controfattuale. Ecco una citazione da un popolare libro di testo sull'inferenza causale:¹⁸⁸

I tribunali statunitensi hanno emanato direttive chiare su ciò che costituisce discriminazione sul lavoro. Secondo i legislatori, "La questione centrale in ogni caso di discriminazione sul lavoro è se il datore di lavoro avrebbe intrapreso la stessa azione se il dipendente fosse stato di razza diversa (età, sesso, religione, origine nazionale, ecc.) e tutto il resto fosse stato lo stesso." (In Carson contro Bethlehem Steel Corp., 70 casi FEP 921, 7th Cir. (1996).)

Purtroppo la situazione non è così semplice. Questa citazione qui invocata – e in molti altri articoli tecnici sull'argomento – esprime l'opinione dei giudici

7° Circuit Court in quel momento. Questa corte è una delle tredici corti d'appello degli Stati Uniti. Il caso ha poco valore precedente; la citazione non può essere considerata una dichiarazione definitiva su cosa significhi discriminazione sul lavoro ai sensi del Titolo VII o della legge sulla parità di protezione.

Più significativo nella giurisprudenza statunitense è lo standard del "ma per la causalità" che ha guadagnato sostegno attraverso una decisione della Corte Suprema degli Stati Uniti del 2020 relativa alla discriminazione sessuale nel caso Bostock contro Clayton County. In riferimento allo statuto del Titolo VII sulla discriminazione sul lavoro contenuto nel Civil Rights Act del 1964, la corte ha affermato:

Sebbene il testo dello statuto non discuta esplicitamente la causalità, è indicativo. La garanzia che ogni persona abbia diritto allo stesso diritto. . . come piace ai cittadini bianchi, dirige la nostra attenzione sul contropartite: cosa sarebbe successo se il querelante fosse stato bianco? Questo focus si adatta naturalmente alla regola ordinaria secondo cui un attore deve provare ma-per causalità.

Sebbene qui appaia il linguaggio dei contropartite, la nozione di causalità ma-per potrebbe non corrispondere effettivamente a un contropartite causale corretto.

Espandendo le modalità di interpretazione del nesso di causalità, la corte ha osservato:

un test "però" ci indirizza a cambiare una cosa alla volta e vedere se il risultato cambia. Se così fosse, abbiamo trovato una causa ma-per.

Cambiare un attributo mantenendo fissi tutti gli altri non è in generale un modo corretto di calcolare i contropartite in un grafico causale. Questa importante questione è stata al centro di un'importante causa contro la discriminazione.

Ammissioni all'università di Harvard

In un processo risalente al 2015, il querelante Students for Fair Admissions (SFFA) sostiene la

discriminazione nelle ammissioni agli studenti universitari di Harvard contro gli asiatici-americani.

Il querelante SFFA è una propaggine di un fondo di difesa legale che mira a porre fine all'uso della razza nelle votazioni, nell'istruzione, negli appalti e nell'occupazione.

Lo studio ha comportato scoperte senza precedenti riguardanti i processi di ammissione e il processo decisionale di ammissione all'istruzione superiore, comprese analisi statistiche dei dati dei candidati a livello individuale degli ultimi cinque cicli di ammissione.

La perizia del querelante di Peter S. Arcidiacono, professore di economia alla Duke University, afferma:

La razza gioca un ruolo significativo nelle decisioni di ammissione. Considera l'esempio di un candidato asiatico-americano che è maschio, non è svantaggiato e ha altre caratteristiche che comportano una probabilità di ammissione del 25%. Cambiare semplicemente la razza del richiedente in bianco – e lasciare invariate tutte le sue altre caratteristiche – aumenterebbe la sua possibilità di ammissione al 36%. Cambiando la sua razza in ispanica (e lasciando tutto

altre caratteristiche sono le stesse) aumenterebbe la sua possibilità di ammissione al 77%. Cambiare la sua razza in afro-americana (ancora una volta, lasciando invariate tutte le altre caratteristiche) aumenterebbe le sue possibilità di ammissione al 95%.

L'accusa del querelante, riassunta sopra, si basa tecnicamente sull'argomentazione secondo cui la parità statistica condizionale non è soddisfatta da un modello delle decisioni di ammissione di Harvard. Il processo decisionale di Harvard non è codificato come regola decisionale formale. Quindi, per parlare formalmente della regola decisionale di Harvard, dobbiamo prima modellare la regola decisionale di Harvard. L'esperto del querelante lo ha fatto adattando un modello di regressione logistica alle precedenti decisioni di ammissione di Harvard in termini di variabili ritenute rilevanti per la decisione di ammissione.

Formalmente, denotando con Y il modello delle decisioni di ammissione di Harvard, con X un insieme di caratteristiche del candidato ritenute rilevanti per l'ammissione, e denotando con A la razza dichiarata dal

richiedente abbiamo che $E[Y | X = x, A = a] < E[Y | \dots, X = x, A = a]$

$X = x, A = a \Rightarrow \exists a$ e qualche valore significativo di $\exists > 0$.

La violazione di questa condizione dipende certamente da quali caratteristiche riteniamo rilevanti per l'ammissione, formalmente, da quali caratteristiche X dovremmo condizionare. In effetti, questo punto è in larga misura la base della risposta dell'esperto dell'imputato David Card, professore di economia all'Università della California, Berkeley. Card sostiene che con una diversa scelta ragionevole di X , che include tra le altre caratteristiche il rendimento del colloquio del richiedente e l'anno in cui ha presentato domanda, la disparità osservata scompare.

La selezione e la discussione di ciò che costituisce le caratteristiche rilevanti è certamente importante per l'interpretazione della parità statistica condizionale. Ma probabilmente una questione più grande è se una violazione della parità statistica condizionale costituisca in primo luogo una prova di discriminazione. Non è solo una questione di aver selezionato le caratteristiche giuste su cui condizionarsi.

Che cosa intende la perizia del querelante con "cambiare la sua razza"? L'interpretazione letterale è quella di "capovolgere" l'attributo razza nell'input del modello senza modificare nessuna delle altre caratteristiche dell'input. Ma un'interpretazione formale in termini di scambio di attributi non è necessariamente ciò che innesca la nostra intuizione morale. Come ora sappiamo, il ribaltamento degli attributi generalmente non produce controllati validi. In effetti, se assumiamo un grafico causale in cui alcune delle caratteristiche rilevanti sono influenzate dalla razza, allora il calcolo dei controllati rispetto alla razza richiederebbe un aggiustamento delle caratteristiche a valle. Cambiare l'attributo razza senza cambiare qualsiasi altro attributo corrisponde a un controllato solo nel caso in cui la razza non abbia nodi discendenti: un presupposto non plausibile.

Al ribaltamento degli attributi viene spesso erroneamente data un'interpretazione causale controllata. Ottenere controllati validi è in generale sostanzialmente più complicato che invertire un singolo attributo indipendentemente dagli altri. In particolare, non possiamo parlare in modo significativo di controllati senza chiarire a cosa esattamente ci riferiamo nel nostro modello causale e come possiamo produrre modelli causali validi. Passiamo a questo importante argomento in seguito.

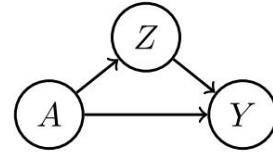


Figura 5.12: La religione come nodo radice.

Validità del modello causale

Consideriamo un'accusa di discriminazione sul lavoro del tipo: le pratiche di assunzione dell'azienda discriminano i candidati di una determinata religione. Supponiamo di voler interrogare questa affermazione utilizzando il meccanismo formale sviluppato in questo capitolo. Ciò richiede innanzitutto di introdurre formalmente un attributo corrispondente all' "appartenenza religiosa" di un individuo.

Il nostro primo tentativo è di modellare l'affiliazione religiosa come un tratto personale o una caratteristica che qualcuno possiede o non possiede. Questa caratteristica, chiamiamola A, può influenzare le scelte relative all'aspetto, alle pratiche sociali e alle variabili rilevanti per il lavoro, come il livello di istruzione Z della persona. Quindi, potremmo iniziare con un modello come il seguente: L'affiliazione

religiosa A è un nodo sorgente in questo grafico, che influenza il livello di istruzione Z della persona . I membri di alcune religioni possono essere allontanati o incoraggiati a ottenere un livello di istruzione più elevato dal loro gruppo sociale di pari. Questa storia è simile a come nel nostro grafico delle ammissioni di Berkeley il sesso influenza la scelta del dipartimento.

Questa visione della religione pone l'onere sulla comprensione dei possibili percorsi indiretti, come A → Z → Y, attraverso i quali la religione può influenzare il risultato.

Potrebbe non esserci una comprensione sufficiente di come un'affiliazione religiosa influisca su numerose altre variabili rilevanti nel corso della vita. Se pensiamo alla religione come a un nodo sorgente in un grafo causale, il suo cambiamento influenzerebbe potenzialmente tutti i nodi a valle. Per ciascuno di questi nodi a valle avremmo bisogno di una chiara comprensione dei meccanismi attraverso i quali la religione influenza il nodo. Da dove verrebbe tale conoscenza scientifica di tali relazioni?

Ma anche la storia causale della religione potrebbe essere diversa. Potrebbe darsi che il conseguimento di un livello di istruzione più elevato induca un individuo a perdere le proprie convinzioni religiose. In effetti, questa scelta di modello è stata avanzata nel lavoro tecnico su questo argomento.³² Empiricamente, i dati del General Social Survey degli Stati Uniti mostrano che la frazione di intervistati che ha cambiato la religione dichiarata almeno una volta durante un periodo di 4 anni variava da circa Dal 20% al 40% circa.¹⁸⁹ Le identità associate alla sessualità e alla classe sociale sono risultate ancora più instabili. Cambiare la propria identità per allinearsi meglio alla propria politica sembrava spiegare in parte questo cambiamento. Da questa prospettiva, l'appartenenza religiosa è influenzata dal livello di istruzione e quindi il grafico potrebbe apparire così:

Questa visione della religione ci obbliga a identificare correttamente le variabili che influenzano l'appartenenza religiosa e sono anche rilevanti per la decisione. Dopotutto, questi sono i

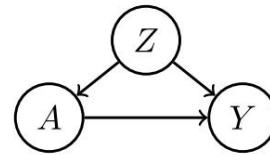


Figura 5.13: La religione come antenato.

confondimenti tra religione e risultato. Forse non è solo il livello di istruzione, ma anche lo status socioeconomico e altri fattori ad avere un'influenza confondente simile .

Ciò che è preoccupante è che nel nostro primo grafico l'istruzione è un mediatore, mentre nel nostro secondo grafico è un confondente. La differenza è importante; per citare Perla:

Come ormai sicuramente saprai, confondere un mediatore con un confondente è uno dei peccati più mortali nell'inferenza causale e può portare all'errore più scandaloso. Quest'ultimo invita all'aggiustamento; il primo lo vieta.¹⁸²

Il punto non è che queste siano le uniche due possibili scelte modellistiche su come l'affiliazione religiosa potrebbe interagire con i processi decisionali. Piuttosto, il punto è che esistono molteplici scelte plausibili. Ciascuna delle nostre scelte di modellazione segue una storia causale naturale. Identificare quale sia giustificato non è un compito facile. Inoltre, non è un compito che possiamo aggirare facendo appello a una sorta di pragmatismo. Diverse scelte di modellazione possono portare ad affermazioni e conseguenze completamente diverse.

Per creare un modello causale valido, dobbiamo fornire chiarezza su quale sia l'oggetto a cui fa riferimento ciascun nodo e quali relazioni esistono tra queste cose. Questo è un problema di ontologia e metafisica. Ma dobbiamo anche conoscere i fatti relativi alle cose a cui facciamo riferimento nei modelli causali. Questo è un problema dell'epistemologia , della teoria della conoscenza.

Questi problemi potrebbero sembrare banali per alcuni oggetti di studio. Potremmo avere forti convinzioni scientificamente giustificate su come interagiscono alcune parti meccaniche di un aereo. Possiamo usare questa conoscenza per diagnosticare in modo affidabile la causa di un incidente aereo. In altri ambiti, soprattutto quelli rilevanti per le controversie sulla discriminazione, la nostra conoscenza dell'argomento è meno stabile e soggetta a dibattito.

Costruzione sociale delle categorie

Le difficoltà che abbiamo incontrato nel nostro esempio motivante sorgono abitualmente quando facciamo affermazioni causali che coinvolgono generi e categorie umane, come razza, religione o genere, e come queste interagiscono con decisioni consequenziali.

Consideriamo il caso della razza. La metafisica della razza è un argomento complesso, molto dibattuto, che oggi presenta una serie di resoconti accademici. Un libro di Glasgow, Haslanger, Jeffers e Spencer rappresenta quattro visioni filosofiche contemporanee su cosa sia la razza.¹⁹⁰ La costruzione delle categorie razziali e la classificazione razziale della razza

individui è inestricabilmente legato a una lunga storia di oppressione, segregazione e pratiche discriminatorie.^{191, 192, 193}

Nella letteratura tecnica sulla discriminazione e la causalità, è comune per i ricercatori modellare la razza come nodo sorgente in un grafico causale, vale a dire che la razza non ha frecce in arrivo. Come nodo di origine può influenzare direttamente e indirettamente una variabile di risultato, ad esempio, ottenere un'offerta di lavoro. Implicita in questa scelta modellistica è una sorta di prospettiva naturalistica che vede la razza come un tratto biologicamente radicato, simile al sesso. Il tratto esiste all'inizio della vita. Altre variabili che arrivano più tardi nella vita, ad esempio l'istruzione e il reddito, diventano così antenati nel grafico causale.

Questa visione della razza ci sfida a identificare tutti i possibili percorsi indiretti attraverso i quali la razza può influenzare il risultato. Ma non è solo questa sfida di modellizzazione che dobbiamo affrontare. La visione della razza come tratto biologicamente fondato è in contrasto con la visione costruttivista sociale della razza.^{194, 195, 196, 190} In questa visione, grosso modo, la razza non ha un forte fondamento biologico ma è piuttosto un costrutto sociale. La razza deriva da una particolare classificazione degli individui da parte della società e dalle esperienze condivise che derivano da tale classificazione. In quanto tale, il sistema sociale circostante di un individuo influenza cos'è la razza e come viene percepita. Nella visione costruttivista, la razza è una categoria socialmente costruita a cui vengono assegnati gli individui.

La sfida con l'adozione di questa visione è che è difficile individuare una serie di nodi che rappresentino fedelmente l'influenza che la società ha sulla razza e sulla percezione della razza. La prospettiva sociale costruttivista non viene fornita con una semplice guida operativa per identificare le strutture causali. In particolare, le categorie costruite socialmente spesso mancano del tipo di modularità richiesta da un diagramma causale. Supponiamo che l'appartenenza al gruppo sia costruita a partire da un insieme di fatti sociali riguardanti il gruppo e dalle pratiche degli individui all'interno del gruppo. Potremmo avere una certa comprensione di come questi fatti e pratiche identificano costitutivamente l'appartenenza al gruppo. Ma potremmo non comprendere come ciascun fattore interagisce individualmente con ogni altro fattore, o se tale scomposizione sia addirittura possibile.¹⁹⁷

Instabilità ontologica

Nonostante le argomentazioni precedenti, i pragmatici potrebbero accusare la nostra discussione di aggiungere complessità non necessaria a quella che ad alcuni potrebbe sembrare una questione di buon senso. Sicuramente potremmo trovare sottigliezze anche in altre caratteristiche, come l'abitudine al fumo o l'esercizio fisico. In che modo la razza è diversa dalle altre cose a cui facciamo riferimento nei modelli causali?

Una differenza importante è una questione di stabilità ontologica. Quando diciamo che la pioggia ha fatto bagnare l'erba ci riferiamo anche ad una comprensione implicita di cosa sia la pioggia, cos'è l'erba e cosa significhi bagnato. Tuttavia, lo troviamo accettabile in questo caso, perché tutte e tre le cose a cui ci riferiamo nella nostra affermazione causale hanno ontologie sufficientemente stabili. Sappiamo a cosa facciamo riferimento quando li invochiamo. A dire il vero, potrebbero esserci delle sottigliezze in ciò che chiamiamo erba. Forse il termine colloquiale erba

non corrisponde ad una categoria botanica precisa, né ad una che è cambiata nel tempo e cambierà ancora in futuro. Tuttavia, facendo l'affermazione causale, affermiamo implicitamente che queste sottigliezze sono irrilevanti per l'affermazione che abbiamo fatto. Sappiamo che l'erba è una pianta e che anche le altre piante si bagnerebbero a causa della pioggia. In breve, crediamo che le ontologie a cui facciamo riferimento siano sufficientemente stabili per le nostre affermazioni .

Non è sempre facile dare un giudizio. Esistono, in generale, almeno due fonti di instabilità ontologica. Uno deriva dal fatto che il mondo cambia nel tempo. Sia il progresso sociale, sia gli eventi politici, sia le nostre stesse attività epistemiche possono rendere obsolete teorie, creare nuove categorie o sconvolgere quelle esistenti.¹⁹⁶ Il lavoro di Hacking descrive un'altra importante fonte di instabilità. Le categorie portano le persone che presumibilmente rientrano in tali categorie a modificare il proprio comportamento in modi possibilmente inaspettati. Gli individui potrebbero conformarsi o disconformarsi alle categorie con cui si confrontano. Di conseguenza, le risposte delle persone, individualmente o collettivamente, invalidano la teoria alla base della categorizzazione. L'hacking lo definisce un "effetto looping".¹⁹⁸ In quanto tale, le categorie sociali stanno spostando gli obiettivi che hanno bisogno revisione costante.

Certificati di stabilità ontologica

Il dibattito sulle categorie umane nei modelli causalì non è affatto nuovo. Ma spesso emerge in una discussione apparentemente non correlata, ma di lunga data, su causalità e manipolazione. Una scuola di pensiero sull'inferenza causale si allinea con il mantra "nessuna causalità senza manipolazione", un punto di vista espresso da Holland in un influente articolo del 1986:

Detto nel modo più schietto e controverso possibile, in questo articolo sostengo la posizione secondo cui le cause sono solo quelle cose che potrebbero, in linea di principio, essere trattamenti negli esperimenti.¹⁹⁹

Holland va oltre sostenendo che le affermazioni che coinvolgono "attributi" lo sono necessariamente dichiarazioni di associazione:

L'unico modo in cui un attributo può cambiare il suo valore è che l'unità cambi in qualche modo e non sia più la stessa unità. Le affermazioni di "causalità" che coinvolgono attributi come "cause" sono sempre affermazioni di associazione tra i valori di un attributo e una variabile di risposta tra le unità di una popolazione.¹⁹⁹

Per fare un esempio, Holland sostiene che la frase "Ha fatto bene all'esame perché è una donna" non significa altro che "il rendimento delle donne all'esame supera, in un certo senso, quello degli uomini".¹⁹⁹

Se credessimo che non esiste causalità senza manipolazione, dovremmo astenerci del tutto dall'includere caratteristiche immutabili nei modelli causali.

Dopotutto, per definizione non esiste alcun meccanismo sperimentale che trasformi gli attributi immutabili in trattamenti.

Il punto di vista di Holland rimane popolare tra i professionisti del modello dei risultati potenziali. Le ipotesi comuni nel modello dei risultati potenziali sono più facili da concettualizzare per analogia con uno studio randomizzato ben progettato. I professionisti in questo contesto sono quindi abituati a concettualizzare le cause come cose che potrebbero, in linea di principio, essere un trattamento in studi randomizzati controllati.

Il desiderio o la necessità di fare affermazioni causali che coinvolgono la razza in un modo o nell'altro non sorge solo nel contesto della discriminazione. Gli epidemiologi incontrano le stesse difficoltà quando affrontano le disparità sanitarie,^{36,200} così come gli scienziati sociali quando ragionano sulla diseguaglianza in termini di povertà, criminalità e istruzione.

I professionisti che si trovano di fronte alla necessità di fare affermazioni causali sulla razza spesso si rivolgono a un particolare trucco concettuale. L'idea è di cambiare oggetto di studio dall'effetto della razza all'effetto della percezione della razza.^{201,202} Ciò a cui tutto questo si riduce è che cambiamo le unità di studio da individui con un attributo razziale a decisorii. Il trattamento diventa esposizione alla razza attraverso alcuni tratti osservabili, come il nome su un CV in un contesto di domanda di lavoro. L'obiettivo dello studio è quindi il modo in cui i decisorii rispondono a tali stimoli razziali nel processo decisionale.

La speranza dietro questa manovra è che l'esposizione alla razza, a differenza della razza stessa, possa essere qualcosa che possiamo controllare, manipolare e sperimentare.

Sebbene questo approccio eviti superficialmente la difficoltà di concettualizzare la manipolazione di caratteristiche immutabili, sposta il peso altrove. Dobbiamo ora risolvere tutti i diversi modi in cui pensiamo che la razza possa essere percepita: attraverso i nomi, il linguaggio, lo stile e ogni sorta di altre caratteristiche e combinazioni degli stessi. Ma non solo. Per fare affermazioni controfattuali, vale a dire l'esposizione alla razza, dovremmo essere in grado di creare le autentiche condizioni di fondo in cui tutte queste caratteristiche percepibili sarebbero emerse in un modo coerente con una diversa categoria razziale. Non c'è modo di costruire accuratamente tali controfattuali senza una chiara comprensione di ciò che intendiamo per categoria di razza.²⁰² Proprio come non possiamo parlare di stregoneria in un modello causale valido per mancanza di qualsiasi base scientifica, non possiamo nemmeno parlare di percezioni di stregoneria in un modello causale valido per la stessa ragione. Allo stesso modo, se ci manca la base ontologica ed epistemica per parlare di razza in un modello causale valido, non è possibile trovare un rimedio facile passando alle percezioni della razza.

In opposizione al punto di vista di Holland, altri studiosi, incluso Pearl, sostengono che la causalità non richiede manipolabilità ma piuttosto una comprensione delle interazioni. Possiamo ragionare su ipotetiche eruzioni vulcaniche senza poter manipolare i vulcani. Possiamo spiegare il meccanismo che provoca le maree senza poter manipolare la luna con alcun intervento fattibile. Ciò che è richiesto è la comprensione dei modi in cui una variabile interagisce con le altre variabili nel modello. Le equazioni strutturali in un modello causale sono funzioni di risposta. Possiamo pensare a un nodo in un grafo causale come se ricevesse messaggi dai suoi nodi principali e rispondesse a tali messaggi. La causalità riguarda quindi chi ascolta chi. Possiamo formare un modello causale una volta che sappiamo come interagiscono i nodi in esso contenuti.

Ma come abbiamo visto, il passaggio concettuale all'interazione – chi ascolta chi – non rende affatto semplice elaborare modelli causali validi. Se i modelli causali organizzano le informazioni scientifiche o empiriche disponibili, inevitabilmente ci sono

limitazioni a quali costrutti possiamo includere in un modello causale senza correre il pericolo di separare il modello dalla realtà. Soprattutto nei sistemi sociotecnici, la conoscenza scientifica potrebbe non essere disponibile in termini di precise funzioni di risposta modulare.

Riteniamo che le cause non debbano essere manipolabili sperimentalmente. Tuttavia, la nostra discussione motiva il fatto che i costrutti a cui si fa riferimento nei modelli causali necessitano di un certificato di stabilità ontologica ed epistemica. La manipolazione può essere interpretata come un approccio un po' pesante per chiarire la natura ontologica di un nodo specificando un meccanismo sperimentale esplicito per manipolare il nodo. Questo è un modo, ma non l'unico, per chiarire a cosa fa riferimento il nodo.

Note del capitolo

Esistono diversi libri di testo introduttivi sul tema della causalità. Per una breve introduzione alla causalità, consultare il manuale di Pearl, Glymour e Jewell,¹⁸⁸ o il libro di testo più completo di Pearl.¹⁸⁰ A livello tecnico, il testo di Pearl enfatizza i grafici causali e i modelli causali strutturali. La nostra esposizione del paradosso di Simpson e dell'Università di Berkeley è stata influenzata dalla discussione di Pearl, aggiornata per un nuovo popolare libro per il pubblico.¹⁸² Tutti questi testi toccano il tema della discriminazione. In questi libri Pearl sostiene che la discriminazione corrisponde all'effetto diretto della categoria sensibile su una decisione.

Il lettore con una mentalità tecnica si divertirà a completare il libro di Pearl con un testo ad accesso libero di Peters, Janzing e Schölkopf¹⁸⁵ disponibile anche online.

Il testo enfatizza due modelli causali variabili e le applicazioni all'apprendimento automatico. Vedi Spirtes, Glymour e Scheines²⁰³ per un'introduzione generale basata sui grafi causali con un'enfasi sulla scoperta dei grafi, cioè sulla deduzione di grafi causali dai dati osservativi.

Morgan e Winship²⁰⁴ si concentrano sulle applicazioni nelle scienze sociali. Imbens e Rubin²⁰⁵ forniscono una panoramica completa del repertorio tecnico dell'inferenza causale nel modello dei risultati potenziali. Angrist e Pischke²⁰⁶ si concentrano sull'inferenza causale e sui potenziali risultati in econometria.

Hernan e Robins²⁰⁷ forniscono un'altra introduzione dettagliata all'inferenza causale che si basa sull'esperienza degli autori in epidemiologia.

Pearl¹⁸⁰ ha già considerato l'esempio della discriminazione di genere nell'ammissione dei laureati alla UC Berkeley di cui abbiamo discusso a lungo. Nella sua discussione, sostiene implicitamente l'idea di discutere la discriminazione basata sui grafici causali esaminando quali percorsi nel grafico vanno dalla variabile sensibile al punto decisionale. L'esempio dell'UC Berkeley è stato discusso in vari altri scritti, come la discussione di Pearl nel Book of Why.¹⁸² Tuttavia, lo sviluppo in questo capitolo differisce significativamente nelle argomentazioni e nelle conclusioni.

Per chiarimenti riguardanti l'interpretazione popolare dell'originale di Simpson articolo,²⁰⁸ vedere l'articolo di Hernan²⁰⁹ e il testo di Pearl.¹⁸⁰

Il tema del ragionamento causale e della discriminazione ha acquisito uno slancio significativo

nella comunità informatica e statistica intorno al 2017. Zhang, Wu e Wu²¹⁰ avevano precedentemente considerato l'analisi della discriminazione tramite effetti causali specifici del percorso. Kusner, Loftus, Russell e Silva¹⁸⁷ hanno introdotto una nozione di equità controfattuale.

Gli autori estendono questa linea di pensiero in un altro lavoro.²¹¹ Chiappa introduce una nozione di equità controfattuale specifica per il percorso.²¹² Kilbertus et al.²¹³ distinguono tra due criteri causali grafici, chiamati discriminazione irrisolta e discriminazione per procura. Entrambe le nozioni corrispondono a percorsi consentiti o non consentiti nei modelli causali. Razieh e Shpitser²¹⁴ concettualizzano la discriminazione come l'influenza dell'attributo sensibile sul risultato lungo determinati percorsi causali non consentiti. Chiappa e Isaac²¹⁵ tengono un tutorial su causalità ed equità con un'enfasi sul dibattito COMPAS. Kasirzadeh e Smart approfondiscono la discussione sulle difficoltà nella costruzione di affermazioni controfattuali causali sulle categorie sociali nel contesto dei problemi di apprendimento automatico.²¹⁶ Esiste anche un'ampia borsa di studio rilevante in

altre discipline che non possiamo esaminare completamente qui. Di rilievo è la vasta letteratura epidemiologica sulle disparità sanitarie. In particolare, gli epidemiologi si sono confrontati con razza e genere in modelli causali. Si veda, ad esempio, l'articolo di VanderWeele e Robinson,²⁰⁰ così come il commento di Krieger all'articolo,²¹⁷ e l'articolo di Krieger sulla discriminazione e le disuguaglianze sanitarie²¹⁸ come punto di partenza.

Abbiamo recuperato i dati sulle ammissioni all'UC Berkeley da <http://www.randomservices.org/random/data/Berkeley.html> il 27 dicembre 2018. C'è qualche discrepanza con i dati visualizzati sulla pagina Wikipedia per il paradosso di Simpson, che non influenzare la nostra discussione.

6

Comprendere la legge sulla discriminazione Unito Stati anti-

In questo capitolo speriamo di darvi un'idea di cosa sia e cosa non sia la legge antidiscriminatoria degli Stati Uniti. Utilizzeremo l'esperienza giuridica statunitense come caso di studio su come regolare la discriminazione. Altri paesi adottano approcci diversi.

Non miriamo a descrivere la legislazione statunitense in modo esaustivo, ma piuttosto a fornire una descrizione stilizzata dei concetti chiave.

Inizieremo con una storia di come sono nate le principali leggi sui diritti civili e trarremo lezioni da questa storia che continuano ad essere rilevanti oggi. Il diritto rappresenta un tentativo di rendere operative le nozioni morali. È importante ed illustrativo. Impareremo dal modo in cui la legge affronta molti compromessi complicati. Ma ne studieremo anche i limiti e spiegheremo perché riteniamo che l'equità algoritmica non debba fermarsi alla conformità legale.

La sezione finale affronta le specifiche della regolamentazione dell'apprendimento automatico. Sebbene la legge antidiscriminazione statunitense sia anteriore all'uso diffuso dell'apprendimento automatico, è altrettanto applicabile se un decisore utilizza l'apprendimento automatico o altre tecniche statistiche. Detto questo, l'apprendimento automatico introduce molte complicazioni nell'applicazione di queste leggi e le leggi esistenti potrebbero essere inadeguate ad affrontare alcuni tipi di discriminazione che si verificano quando è coinvolto l'apprendimento automatico. Allo stesso tempo, crediamo che ci sia anche l'opportunità di esercitare nuovi strumenti normativi per frenare la discriminazione algoritmica.

Questo capitolo può essere saltato in una prima lettura del libro, ma vale la pena sottolineare alcuni collegamenti. La prima sezione approfondisce un punto di vista centrale del libro, in particolare il capitolo 4, ovvero che attributi come la razza e il genere sono salienti perché storicamente sono serviti come principi organizzativi di molte società. Quella sezione istituisce anche il capitolo 8 che concepisce la discriminazione in modo più ampio rispetto ai momenti discreti del processo decisionale. La sezione sui limiti della legge motiva un altro tema centrale di questo libro, che utilizza i dibattiti sull'apprendimento automatico e sulla discriminazione come un'opportunità per rivisitare i fondamenti morali dell'equità.

Storia e panoramica della legislazione antidiscriminatoria statunitense

Ogni centimetro di tutela dei diritti civili incorporato nella legge è stato combattuto e conquistato duramente attraverso decenni di attivismo. In questa sezione descriviamo brevemente queste storie di oppressione e discriminazione, i movimenti che sorse in risposta ad esse e i cambiamenti legali che apportarono.

Diritti civili dei neri

Il movimento per i diritti civili dei neri, spesso chiamato semplicemente movimento per i diritti civili, affonda le sue radici nella schiavitù negli Stati Uniti e nella dilagante discriminazione razziale che persistette dopo la sua abolizione. Il periodo immediatamente successivo alla guerra civile americana e all'abolizione della schiavitù (all'incirca dal 1865 al 1877) è chiamato l'era della ricostruzione. Il risultato è stato un progresso sostanziale nel campo dei diritti civili. In particolare, la Costituzione è stata modificata per abolire la schiavitù (13° emendamento), richiedere pari protezione ai sensi delle leggi (14° emendamento) e garantire il diritto di voto indipendentemente dalla razza (15° emendamento).

Tuttavia, questi progressi furono rapidamente annullati quando i suprematisti bianchi acquisirono il controllo politico negli stati del sud, inaugurando la cosiddetta era di Jim Crow, un periodo di circa 75 anni in cui lo stato orchestrò una dura segregazione razziale, discriminazione e privazione quasi totale dei diritti civili dei neri. Quasi ogni aspetto della vita era sottoposto a segregazione razziale, compresi i quartieri residenziali, le scuole, i luoghi di lavoro e i luoghi di accoglienza pubblica come ristoranti e alberghi.

Questa segregazione fu benedetta dalla Corte Suprema nel 1896, quando stabilì che le leggi che imponevano la segregazione non violavano la clausola di uguaglianza di protezione prevista dalla dottrina "separati ma uguali".²¹⁹ Ma in pratica, le cose erano tutt'altro che uguali. I lavori a disposizione dei neri erano solitamente pagati molto meno, le scuole erano sottoponenziate e soggette a chiusura, e gli alloggi erano meno numerosi e di qualità inferiore. Ancora negli anni '50, un viaggio attraverso il paese da parte di una persona di colore avrebbe comportato un grande pericolo, ad esempio presentarsi di notte in una piccola città e vedersi rifiutare un posto dove stare.²²⁰ I neri non potevano sfidare democraticamente queste leggi in quanto gli stati hanno eretto numerose barriere pratiche al voto – apparentemente neutrali rispetto alla razza, ma con effetti molto diversi a seconda della razza – e i neri alle urne sono stati spesso accolti con violenza. Di conseguenza, la privazione dei diritti civili è stata molto efficace. Ad esempio, in Louisiana fino alla metà degli anni '40, meno dell'1% degli afroamericani era registrato per votare.²²¹ (I limiti dei dati precludono una valutazione a livello nazionale dell'efficacia della privazione dei diritti civili).

Nel frattempo, negli stati del Nord, la discriminazione razziale operava in modi più indiretti. Le leggi sulla zonizzazione residenziale che proibivano alloggi ad alta densità e a basso costo furono utilizzate per tenere i residenti neri più poveri fuori dai quartieri bianchi. La pratica del "redlining" da parte delle banche, orchestrata in una certa misura dai regolatori federali, ha limitato la disponibilità di credito, soprattutto di mutui, in quartieri specifici.²²² La giustificazione addotta era il livello di rischio, ma ha avuto l'effetto di discriminare le comunità nere. Un'altra tecnica prevalente per ottenere la segregazione era l'uso di alleanze razzialmente restrittive in cui i proprietari di immobili in un quartiere

stipulò un contratto per non vendere o affittare a persone non

bianche.¹ Il movimento per i diritti civili emerse tra la fine del 1800 e l'inizio del 1900 per affrontare queste diffuse pratiche di razzismo. In generale, il movimento adottò due strategie complementari: una era quella di sfidare le leggi ingiuste e l'altra era quella di far avanzare la società nera entro i limiti della segregazione e della discriminazione. Un momento chiave nel primo polo fu la formazione dell'Associazione Nazionale per il Progresso delle Persone di Colore nel 1909. Oltre a fare lobbying e a intentare causa contro le leggi Jim Crow, cercò di lottare contro il linciaggio. Importanti sforzi nell'ambito del secondo polo includevano il movimento imprenditoriale nero – il 1900-1930 è stato definito l'età dell'oro delle imprese nere²²³ – e notevoli risultati nel campo dell'istruzione. Molti dei college e delle università storicamente neri furono fondati durante l'era di Jim Crow.

Dopo decenni di attivismo, un momento epocale fu la sentenza della Corte Suprema del 1954 che dichiarò incostituzionale la segregazione delle scuole pubbliche. Ciò diede inizio al graduale smantellamento del sistema Jim Crow, un processo che avrebbe richiesto decenni e i cui effetti avvertiamo ancora oggi. Le vittorie in tribunale hanno ulteriormente galvanizzato il movimento, portando a un attivismo più intenso e a proteste di massa. Ciò portò a importanti leggi federali nel decennio successivo: il Civil Rights Act del 1960 e il Voting Rights Act del 1965, entrambi mirati alla repressione degli elettori, e il Civil Rights Act del 1964 e il Fair Housing Act del 1968 che prendevano di mira i privati. discriminazione. Discuteremo gli ultimi due in dettaglio nel corso di questo capitolo.

Le leggi contro la discriminazione erano chiaramente un prodotto della storia e di tendenze decennali o secolari: la schiavitù, Jim Crow e il movimento per i diritti civili. Allo stesso tempo, le loro cause prossime erano spesso eventi specifici e imprevedibili. Ad esempio, l'assassinio di Martin Luther King Jr. ha dato impulso all'approvazione del Fair Housing Act. Riflettono anche i compromessi politici necessari per garantire il loro passaggio. Ad esempio, il titolo VII del Civil Rights Act del 1964 ha creato la Commissione per le pari opportunità di lavoro; è stato privato dei poteri esecutivi che erano presenti nella formulazione originaria del titolo.²²⁴

Discriminazione di genere

La lotta per l'uguaglianza di genere ha anche una lunga e storica storia di attivismo.

Nel 1800, la legge non riconosceva i diritti fondamentali delle donne, compreso il voto e il possesso di proprietà. Cambiare questa situazione era l'obiettivo principale delle femministe della prima ondata, le cui strategie includevano sostegno, disobbedienza civile, lobbying e azioni legali. Il momento culminante fu la ratifica del 19° emendamento nel 1920, che garantiva alle donne il diritto di voto (tuttavia, come discusso sopra, il diritto di voto delle donne nere era ancora limitato nel Sud). Il femminismo della seconda ondata iniziò negli anni '60.

Ha preso di mira gli stereotipi sul ruolo delle donne nella società, la discriminazione privata nell'istruzione e nel lavoro e i diritti fisici, compresi i diritti riproduttivi e la violenza domestica.

Nei primi anni del dopoguerra, in alcuni casi le norme di genere regredirono

¹Questa pratica continuò fino a quando la Corte Suprema la annullò nel 1948 (*Shelley v. Kraemer*), sostenendo che, anche se si trattava di contratti privati, se lo Stato dovesse applicarli, violerebbe la clausola costituzionale di pari protezione.

modi (ad esempio, le donne persero l'accesso ai posti di lavoro che erano stati disponibili per loro a causa della guerra) che furono probabilmente un impulso per il movimento.²²⁵ Due delle prime vittorie legislative furono l'Equal Pay Act del 1963 e il Titolo VII del Civil Rights Act del 1963. 1964 che proibiva la discriminazione sul lavoro.²

Tuttavia, a causa della mancata applicazione di cui sopra, inizialmente questi non ebbero un grande impatto e il movimento non fece altro che intensificarsi. Una pietra miliare importante fu la fondazione dell'Organizzazione Nazionale delle Donne nel 1965. Prendendo in prestito le strategie dal movimento per i diritti civili dei neri, le femministe della seconda ondata adottarono un piano per ricorrere in tribunale per garantire la protezione delle donne. Una notevole vittoria della corte nel decennio successivo fu l'espansione del diritto all'aborto da parte della Corte Suprema.²²⁶ Sul fronte legislativo, due importanti risultati per l'uguaglianza di genere furono il Titolo IX dell'Education Amendments Act del 1972 che proibiva la discriminazione sessuale in programmi educativi finanziati a livello federale e l'Equal Credit Opportunity Act che proibiva la discriminazione sessuale nel credito.

L'istruzione, in particolare l'istruzione superiore, e il credito erano entrambi settori importanti per i diritti delle donne. Storicamente, molte università d'élite semplicemente non accettavano le donne. Anche negli anni '70, le donne dovevano affrontare molti ostacoli nel mondo accademico: molestie sessuali, ostacoli più elevati all'ammissione, totale esclusione da alcuni campi di alto livello come diritto e medicina e limitate opportunità sportive. Allo stesso modo, anche la discriminazione creditizia negli anni '70 fu grave, come ad esempio richiedere alle donne di richiedere nuovamente il credito al momento del matrimonio, di solito a nome del marito.²²⁷ Dopo questo periodo, l'attenzione del movimento femminista si espansero oltre le principali vittorie legislative per includere la messa in discussione del genere come costrutto sociale.

Diritti civili LGBTQ

Le leggi discriminatorie contro le persone LGBTQ sono state storicamente numerose: divieto di alcuni comportamenti sessuali (ad esempio, leggi anti-sodoma²²⁸), mancanza di diritti matrimoniai, divieti di servizio militare e di alcune altre posizioni governative, incapacità di proibire la discriminazione privata e di trattare i crimini d'odio come tale, e persino il divieto della letteratura che difende i diritti dei gay ai sensi delle leggi sull'oscenità.

L'attivismo provvisorio iniziò negli anni '50 con le prime modifiche legali arrivate all'inizio degli anni '60. Un movimento cruciale furono le rivolte di Stonewall del 1969, una serie di proteste in risposta a un raid della polizia in un bar gay di New York. Le conseguenze di questo evento hanno dato il via alla spinta per i diritti LGBTQ negli Stati Uniti, compreso il movimento del gay pride per ottenere visibilità e accettazione. Nel 1973, il Manuale diagnostico e statistico dei disturbi mentali dell'American Psychiatric Association abbandonò l'omosessualità come disturbo, segnalando (e favorendo) un importante cambiamento negli atteggiamenti. L'elenco delle modifiche legislative è lungo e in corso. Includono modifiche state per stato alle leggi che riguardano la sodomia, l'uguaglianza dei matrimoni, la discriminazione privata e i crimini d'odio; una decisione della Corte Suprema del 2003 che dichiara incostituzionali le leggi anti-sodoma;²²⁹ e una decisione della Corte Suprema del 2015 che garantisce il diritto di sposarsi per le coppie dello stesso sesso a livello nazionale.²³⁰ Parallelamente, la spinta per i diritti LGBTQ nel settore privato è in parte progredi-

²²⁸Sebbene quest'ultima fosse principalmente una risposta al movimento per i diritti civili dei neri, il sesso è stato aggiunto come categoria protetta in un emendamento dell'ultimo minuto.

interpretando i divieti statutari esistenti sulla discriminazione sessuale, come il Titolo VII del Civil Rights Act del 1964, per comprendere la discriminazione basata sull'orientamento sessuale e sull'identità di genere.²³¹

Leggi sulla disabilità

Un'altra dimensione dell'identità coperta dagli statuti antidiscriminazione è la disabilità. Oggi, negli Stati Uniti, oltre un quarto degli adulti soffre di qualche tipo di disabilità, tra cui disabilità motorie, cecità o altre disabilità visive, sordità o altre disabilità uditive e disabilità cognitive.²³² Queste e altre disabilità sono identità distinte corrispondenti a diverse esperienze vissute e, a volte, culture.³ Tuttavia, l'emergere di una coalizione e di un'identità trasversale ha consentito una difesa più efficace dei diritti dei disabili. Questo movimento prese piede nei decenni successivi alla seconda guerra mondiale. Gli attivisti miravano a rendere la disabilità visibile, piuttosto che stigmatizzata, compatita e nascosta, e cercavano di raggiungere una vita indipendente. Come altri movimenti per i diritti, le persone disabili si sono trovate ad affrontare molteplici barriere che si rafforzavano a vicenda: l'atteggiamento della società nei confronti della disabilità e dei disabili, la mancanza di strutture fisiche e di tecnologie assistive, e politiche discriminatorie.²³⁴ Gli atteggiamenti che hanno frenato le persone disabili non erano solo pregiudizi, ma anche visioni errate della disabilità come residente nella persona (il modello medico) invece che, o in aggiunta a, creata dalle barriere presenti nella società (il modello sociale). La prima legge federale a tutela dei diritti dei disabili è stata il Rehabilitation Act del 1973 che proibiva la discriminazione dei disabili nei programmi finanziati a livello federale. L'attivismo verso un ampio statuto sui diritti civili continuò, prendendo a modello il Civil Rights Act del 1964. Questi sforzi culminarono nell'Americans with Disabilities Act (ADA) del 1990. Sebbene l'ADA presenti molte somiglianze con gli altri statuti sui diritti civili, presenta anche importanti differenze dovute alla sua enfasi sull'alloggio oltre alla non discriminazione formale.

Lezioni

Le storie dei vari movimenti per i diritti civili contengono diverse lezioni che continuano ad essere rilevanti oggi. Innanzitutto, la legge è uno strumento politico: può essere utilizzata per discriminare, per creare le condizioni in cui la discriminazione possa prosperare, o per sfidare la discriminazione. Può essere uno strumento di sottomissione o di liberazione.

Le leggi possono essere esteriormente neutre, ma sono create, interpretate e applicate da attori che rispondono ai cambiamenti dei tempi e all'attivismo. Le decisioni dei tribunali sono influenzate anche dall'attivismo contemporaneo e persino dagli studi accademici.

La nostra breve discussione storica aiuta anche a spiegare perché alcuni settori sono regolamentati e non altri. Istruzione, occupazione, alloggio, credito e alloggi pubblici sono ambiti estremamente importanti per la vita delle persone e hanno avuto storie di discriminazione utilizzate deliberatamente per subordinare alcuni gruppi.⁴

³In particolare la cultura dei sordi; per un'introduzione vedere 233

⁴ Inoltre, esistono limitazioni costituzionali alla capacità del Congresso di regolamentare il settore privato

Una conseguenza di questo approccio settoriale è che la legge può essere adattata alle particolarità del settore nel tentativo di evitare scappatoie. Ad esempio, il Fair Housing Act comprende l'intera gamma di pratiche relative all'edilizia abitativa, comprese vendite, affitti, pubblicità e finanziamenti. Elenca (e vieta) vari modi in cui gli agenti immobiliari possono fuorviare o scoraggiare subdolamente i clienti appartenenti a classi protette. Riconoscendo l'importanza del finanziamento per garantire l'edilizia abitativa, vieta la discriminazione nel finanziamento rispetto a "acquisto, costruzione, miglioramento, riparazione o manutenzione". Vieta perfino la pubblicità che indica una preferenza discriminatoria. E ciò include non solo affermazioni categoriche come "niente bambini", ma anche il targeting degli annunci pubblicitari in determinate regioni geografiche in modo correlato alla razza e alla selezione degli attori utilizzati nella pubblicità.

In molti casi questi tentativi di evitare scappatoie hanno retto bene di fronte ai recenti sviluppi tecnologici. Il divieto di pubblicità discriminatoria ha costretto le piattaforme pubblicitarie online a evitare il targeting discriminatorio degli annunci immobiliari.²³⁵ Ma non è sempre così. Le piattaforme di ride hailing sono in grado di eludere la responsabilità prevista dal Titolo VII (discriminazione sul lavoro) anche se licenziano i conducenti sulla base delle valutazioni (potenzialmente discriminatorie) fornite dai passeggeri.²³⁶

Anche se le leggi sono specifiche per settore, è difficile comprendere la discriminazione guardando a qualsiasi insieme di istituzioni (come l'occupazione o l'istruzione, tanto meno una singola organizzazione) in modo isolato. La storia ci mostra che tendono ad esserci molteplici sistemi di oppressione interconnessi che operano in tandem, come la politica immobiliare federale e la discriminazione nel settore privato. Allo stesso modo, il confine tra discriminazione statale e privata non è sempre chiaro.

La storia mostra anche che, quando vengono interrotte, le gerarchie tendono a riaffermarsi con altri mezzi. Ad esempio, la fine della segregazione de jure ha accelerato il fenomeno della "fuga dei bianchi" dalle città alle periferie, esacerbando la segregazione di fatto. Non solo il progresso è discontinuo, ma la regressione è possibile. Ad esempio, Woodrow Wilson e la sua amministrazione segregarono gran parte della forza lavoro federale negli anni '10, erodendo alcuni dei guadagni ottenuti dai neri nei decenni precedenti. E mentre stavamo scrivendo questo capitolo, la Corte Suprema ha annullato Roe v. Wade, ponendo fine alla protezione federale del diritto all'aborto e consentendo severe restrizioni all'aborto in molti stati.

Un altro punto importante che non emerge dalle leggi stesse è che le varie dimensioni protette dell'identità hanno storie complesse e distinte di discriminazione e attivismo, anche se gli statuti tentano di trattarle tutte in modo uniforme e formale. Anche all'interno di un'unica dimensione come quella etnica, l'oppressione e le lotte dei diversi gruppi assumono forme drasticamente diverse. I nativi americani hanno sopportato un secolo di tentativi di assimilazione forzata in cui i bambini venivano mandati in collegi e veniva loro chiesto di abbandonare la loro cultura. Il Chinese Exclusion Act del 1882 eliminò del tutto l'immigrazione dei cinesi per oltre mezzo secolo e rese le condizioni inospitali per la comunità di immigrati cinesi già esistente. Durante la seconda guerra mondiale, oltre 100.000 persone di origine giapponese, la maggior parte delle quali erano cittadini statunitensi, furono internate nei campi di concentramento

discriminazione.

con la scusa che fossero sleali verso il paese. Questi sono solo alcuni degli episodi più raccapriccianti di discriminazione sulla base della razza, dell'etnia e dell'origine nazionale nella storia degli Stati Uniti, concentrati sulle azioni del governo. La discriminazione basata sull'origine nazionale era spesso una forma sottilmente velata di discriminazione razziale. Pertanto, sebbene l'elenco degli attributi protetti dalla legge possa aumentare nel tempo, non è arbitrario ed è profondamente influenzato

dalla storia.⁵ L'uguaglianza ai sensi della legge rimane una nozione contestata e in evoluzione. Ciò è particolarmente vero quando l'antidiscriminazione si scontra con qualche valore o principio compensativo, come la libertà religiosa o la limitazione dell'autorità statale. E poiché la legge è intrecciata con le nostre vite e i nostri mezzi di sostentamento in tanti modi, l'uguaglianza davanti alla legge, in senso lato, richiede molto di più della non discriminazione formale. Considera l'uguaglianza di genere. Il ventaglio degli interventi giuridici necessari per realizzarlo è lungo e in crescita. Oltre al diritto di voto e al divieto della discriminazione sessuale e di genere, include il divieto della discriminazione sulla gravidanza e sullo stato civile, il contenimento delle molestie e della violenza sessuale, il diritto all'aborto, le leggi sul congedo di maternità e i sussidi per l'assistenza all'infanzia. Ognuna di queste battaglie ha molti fronti. Ad esempio, il movimento #MeToo ha portato alla luce il ruolo delle clausole di non denigrazione da parte dei datori di lavoro negli accordi volti a mettere a tacere le vittime di molestie sessuali sul posto di lavoro, ed è in corso uno sforzo per vietare tali clausole.

Infine, il cambiamento legislativo non rappresenta la fine del percorso ma, in un certo senso, l'inizio. Gli effetti delle discriminazioni passate tendono a lasciare un'impronta duratura. La legge stessa, date le realtà politiche, non può fare molto per cancellare gli effetti di quella storia.

Tabella 6.1: Una sintesi delle principali leggi antidiscriminazione : titoli VI e VII del Civil Rights Act del 1964, il Fair Housing Act, il titolo IX dell'Education Amendments Act del 1972, l'Equal Credit Opportunity Act e gli americani con Legge sulle disabilità.

Legge	Anno	Entità coperte e attività regolamentate	Categorie protette (* = aggiunte successivamente)
Titolo VII	1964	Datori di lavoro, agenzie per l'impiego, sindacati	Razza, colore, religione, sesso, origine nazionale, gravidanza*
Titolo VI 1964	Qualsiasi organizzazione che riceve finanziamenti federali (a causa dell'ampiezza, non elenca le attività regolamentate)	vendite, affitti, pubblicità e finanziamento di alloggi	Razza, colore, origine nazionale
FHA 1968			Razza, colore, religione, origine nazionale, sesso(*?), handicap, stato familiare.

⁵ Nel contesto della dottrina dell'uguaglianza di protezione, la Corte Suprema ha esplicitamente elencato i criteri che qualificano una caratteristica per la protezione ("controllo rafforzato"): una storia di discriminazione passata, impotenza politica, irrilevanza di una caratteristica per la capacità di un individuo di contribuire o partecipare alla società, e immutabilità.²³⁷

Legge	Anno	Entità coperte e attività regolamentate	Categorie protette (* = aggiunte successivamente)
Titolo IX 1972		Programmi educativi che ricevono finanziamenti federali: assunzioni, retribuzione, grado, molestie sessuali, ritorsioni, segregazione ed educazione	Sesso
ECOA 1974		omosessuale Creditori (banche, piccole società di prestito e finanziarie, negozi al dettaglio e grandi magazzini, società di carte di credito e cooperative di credito).	Razza*, sesso, età*, nazionalità*, stato civile, ricezione del pubblico assistenza*
ADA 1990		Datori di lavoro, servizi pubblici, alloggi pubblici	disabilità; certificato di invalidità; percezione della disabilità

Alcuni fondamenti del sistema giuridico americano

La Costituzione degli Stati Uniti è la legge ultima del paese. La costituzione ha creato i tre rami del governo: il legislativo (Congresso), l'esecutivo (il presidente, le agenzie esecutive e altri che riferiscono al presidente) e il giudiziario (la Corte Suprema e altri tribunali). Tutti e tre i rami hanno ruoli importanti quando si tratta di leggi antidiscriminatorie. Anche i governi e le leggi statali e locali svolgono un ruolo importante nell'antidiscriminazione, ma ne parleremo meno a causa della nostra attenzione pedagogica alla legge federale.

Prima di arrivare ai tre rami, vale la pena notare che la Costituzione stessa contiene due elementi rilevanti per la legge sulla discriminazione: il diritto al giusto processo legale (quinto e quattordicesimo emendamento) e il diritto a pari protezione ai sensi delle leggi (quattordicesimo emendamento). Entrambi questi fattori limitano la capacità del governo di discriminare. La legge sulla parità di protezione è stata talvolta utilizzata anche per ridurre la discriminazione privata. Il giusto processo è stato sollevato come difesa dagli imputati di cause legali per discriminazione sostenendo che le leggi che limitano la loro capacità di discriminare violano i loro diritti al giusto processo.

Il ruolo del Congresso (ramo legislativo)

Le leggi approvate dal Congresso sono chiamate leggi statutarie, in contrapposizione alla legge costituzionale, alla giurisprudenza e ad altri tipi di legge. Abbiamo già incontrato alcune delle principali leggi antidiscriminazione. Ma ci sono molte barriere pratiche e politiche all'azione del Congresso, e statuti o emendamenti sono relativamente rari. Pertanto, per rimanere rilevanti in un mondo in cambiamento, le leggi sono generalmente politiche formulate in modo ampio e non tentano di anticipare le sfumature di ogni situazione in cui potrebbero essere applicate. Per interpretare e applicare queste politiche, il Congresso delega l'autorità ad agenzie federali come la Commissione per le pari opportunità di lavoro (istituita dal Civil Rights Act del 1964). Anche i tribunali svolgono un ruolo interpretativo fondamentale

funzione, oltre a mantenere un controllo sul potere del Congresso stesso.

Ci sono tre principali poteri legislativi che il Congresso ha utilizzato per emanare statuti antidiscriminazione entro i limiti della sua autorità costituzionale. La prima è la clausola commerciale, che consente al Congresso di regolamentare il commercio interstatale. Il significato di questa clausola è stato interpretato estensivamente dalla Corte Suprema.⁶ La normativa in materia di antidiscriminazione nel lavoro e nel credito trova fondamento nella Commerce Clause.⁷ Il secondo potere deriva dal Quattordicesimo Emendamento, che garantisce a tutti i cittadini l'eguale tutela delle leggi, e conferisce inoltre al Congresso il potere di applicarlo attraverso una legislazione adeguata. Sebbene la portata del coinvolgimento statale necessario affinché il Quattordicesimo Emendamento venga applicato non è ancora definita, il Fair Housing Act e l'Americans with Disabilities Act devono in parte la loro base costituzionale a questo potere. Infine, il Congresso ha il "potere della borsa": la capacità di attuare obiettivi politici controllando la spesa e minacciando di trattenere i finanziamenti federali per gli enti che non riescono a soddisfare determinati obblighi. Il Titolo VI del Civil Rights Act del 1964 e il Titolo IX dell'Educational Amendments Act del 1972 rientrano in questa categoria, motivo per cui coprivano solo le organizzazioni che ricevono finanziamenti federali.

Il Congresso ha usato il suo potere per emanare statuti contro la discriminazione che coprono un'ampia gamma di attività. Tuttavia, ci sono molte lacune e limitazioni nella legge federale antidiscriminazione, in parte a causa di limitazioni costituzionali e in parte perché il Congresso non è riuscito ad agire. Di conseguenza, le leggi statali talvolta colmano queste lacune.

Il ruolo dei tribunali (ramo giudiziario)

Gli Stati Uniti adottano un sistema di common law, il che significa che i tribunali hanno il potere di emanare leggi che guidano le decisioni nei casi futuri. Questo è il concetto di precedente. Nelle controversie in cui i fatti o i principi sono simili a casi precedenti decisi dai tribunali competenti, i giudici sono tenuti a seguire il ragionamento utilizzato nella decisione passata (il precedente). Allo stesso modo, i tribunali hanno il compito di interpretare le leggi statutarie e la Costituzione. Questo insieme di precedenti viene definito giurisprudenza e può essere vincolante come qualsiasi altra legge. Ad esempio, l'importante concetto di impatto disparato, in base al quale le pratiche decisionali possono essere illegali se hanno effetti sproporzionali anche se visivamente neutre e senza intenti discriminatori, è il risultato di una decisione della Corte Suprema che interpreta la portata di una legge statutaria. La maggior parte dell'Europa, al contrario, adotta un sistema di diritto civile, il che significa che la legislazione è la fonte primaria del diritto e le decisioni giudiziarie hanno meno valore come precedente.

L'organizzazione gerarchica dei tribunali determina quali precedenti sono vincolanti per una particolare controversia. I tribunali federali sono organizzati in tre livelli: l'

⁶Tra il 1937 e il 1995, un periodo che comprende tutte le leggi discusse sopra, non una sola legge (riguardante o meno la discriminazione) fu invalidata dalla Corte sulla base del fatto che il Congresso eccedeva i suoi poteri ai sensi della clausola commerciale.

⁷Quando fu promulgato il Civil Rights Act del 1964, il suo Titolo II, che proibisce la discriminazione negli alloggi pubblici, fu notoriamente contestato come incostituzionale dal proprietario di un motel ad Atlanta, in Georgia. La Corte Suprema ha confermato la costituzionalità del Titolo II, in un caso che da allora ha avuto un enorme valore precedente.²³⁸

tribunali distrettuali in basso, tredici corti d'appello (note anche come tribunali di circoscrizione) sopra di loro e la Corte Suprema in alto. Le decisioni della Corte Suprema sono vincolanti per tutti i tribunali di grado inferiore e le decisioni delle corti d'appello sono vincolanti per i corrispondenti tribunali distrettuali.⁸ Le corti d'appello esaminano solo casi in appello, cioè quando una delle parti adduce un errore materiale nella decisione di un tribunale distrettuale. La Corte Suprema, a sua volta, di solito esamina solo i ricorsi contro le decisioni dei tribunali circoscrizionali. La Corte Suprema non è tenuta ad accettare istanze di revisione; infatti, concede la revisione solo in una piccola parte delle richieste.

Nell'interpretare le leggi, i tribunali adottano metodi sia testuali che contestuali. Il primo si limita al significato chiaro della legge stessa, mentre il secondo guarda a fonti esterne al testo della legge, come l'intento originariamente espresso dai legislatori e lo scopo della legge. L'importanza dei fattori contestuali è un argomento controverso e i giudici differiscono nei loro approcci.

Finora abbiamo parlato del ruolo dei tribunali nel legiferare. Naturalmente la funzione primaria dei tribunali è quella di giudicare i singoli casi. Sono quindi necessarie alcune note sulla procedura giudiziaria. Il contenzioso può essere civile o penale. Le cause civili riguardano torti contro privati; la maggior parte delle controversie legate alla discriminazione rientrano in questa categoria, con poche eccezioni come i crimini d'odio. I casi penali comportano violazioni del diritto penale e possono essere intentati solo dal governo.⁹

Una caratteristica centrale della procedura giudiziaria statunitense è il ricorso al contraddittorio. Le due parti in causa sono l'attore (che sporge denuncia sostenendo di aver subito un torto) e il convenuto (che si presume abbia commesso il torto). Entrambi sono tipicamente rappresentati da avvocati, che hanno molto potere nel determinare come si svolge il caso, con il giudice che ha un ruolo relativamente passivo come arbitro e non inquisitore.

Un esempio riunirà gli aspetti del sistema giudiziario di cui abbiamo discusso finora. Consideriamo la questione se i siti web di ristoranti, rivenditori, ecc. debbano essere accessibili alle persone non vedenti. L'Americans with Disabilities Act vieta di escludere le persone con disabilità dall'utilizzo dei servizi di un luogo di alloggio pubblico. Ma questo include anche i siti web?

Il Congresso non avrebbe potuto prevedere questa questione nel 1991, quindi lo statuto (nonostante sia insolitamente dettagliato, contando oltre 20.000 parole) non affronta direttamente questa questione. Un gruppo di tribunali circoscrizionali ha esaminato l'intento e lo scopo del Congresso e ha scoperto che i siti web stessi possono essere considerati luoghi di accoglienza pubblica in linea con l'"ampio mandato", lo "scopo ampio" e il "carattere globale" dell'ADA. Un altro gruppo di tribunali circoscrizionali ha adottato un approccio più testuale e ritiene fondamentale che lo statuto si applichi ai servizi di un luogo di alloggio pubblico, non ai servizi in un luogo di alloggio pubblico. Così,

⁸Tutti questi fanno parte del sistema dei tribunali federali. I tribunali statali sono separati e di fatto esaminano la grande maggioranza dei casi, ma i tribunali federali hanno giurisdizione sulle controversie che coinvolgono la legge federale e sono quindi di maggiore rilevanza per noi.

⁹Che cosa renda un atto un crimine piuttosto che un reato civile (oltre a essere giuridicamente classificato come tale) è una questione profonda. Una differenza è la gravità percepita del torto, meritevole di punizione contro l'autore del reato e non solo contro la vittima che viene guarita. Un altro è la natura della parte lesa. I crimini possono essere considerati come un reato contro lo Stato o contro la società, che lo Stato ha interesse a prevenire per evitare il crollo dell'ordine sociale.

finché esiste un “nesso” sufficiente tra il luogo fisico e il sito web, il requisito di accessibilità si estende al sito web. Nell’aprile 2021, un tribunale di circoscrizione si è pronunciato diversamente, interpretando il testo dello statuto nel senso che solo i luoghi fisici possono essere luoghi di alloggio pubblico e respingendo anche lo standard del “nexus” adottato dal secondo gruppo di tribunali. Quando i tribunali circoscrizionali sono divisi in questo modo, di solito è necessario che intervenga la Corte Suprema per risolvere l’incoerenza, ma ciò potrebbe richiedere molti anni.

Uno dei motivi di questo stato di cose è che il Dipartimento di Giustizia, che ha il compito di emanare regolamenti per attuare l’ADA, non ha emesso un regolamento finale sul fatto se i siti web siano luoghi di accoglienza pubblica e, in caso affermativo, quale standard di accessibilità essi bisognerebbe soddisfare. Pertanto i tribunali hanno dovuto esercitare un maggiore grado di discrezionalità interpretativa rispetto a quanto avrebbero altrimenti fatto. In alcuni tribunali ha inoltre suscitato la preoccupazione che imporre requisiti di accessibilità senza stabilire uno standard chiaro priverebbe gli imputati del loro diritto costituzionale a un giusto processo. Ciò evidenzia l’importanza dei dipartimenti esecutivi e delle agenzie federali, di cui parleremo in seguito.

Il ruolo delle agenzie federali (ramo esecutivo)

Le principali funzioni antidiscriminazione delle agenzie federali sono la regolamentazione, l’orientamento e l’applicazione della legge. Ad esempio, l’Equal Credit Opportunity Act rende in generale illegale la discriminazione creditizia, ma lascia alla Federal Reserve il compito di redigere e interpretare i regolamenti che implementano questo mandato (il Congresso ha successivamente trasferito questa autorità al Consumer Financial Protection Bureau). Questo processo è chiamato regolamentazione. Le norme che ne derivano costituiscono diritto amministrativo e hanno forza di legge accanto alla legge statutaria e alla giurisprudenza.

Le regole differiscono leggermente dalle linee guida. Un gruppo di agenzie guidate dalla Commissione per le pari opportunità di lavoro ha pubblicato nel 1978 le Linee guida uniformi per le procedure di selezione dei dipendenti che definiscono un quadro per garantire che i test e le altre procedure di selezione dei dipendenti siano conformi al VII del Civil Rights Act del 1964. sono ampiamente utilizzati dai datori di lavoro. Ma le Linee Guida Uniformi non costituiscono legge. Ad essi si fa spesso riferimento nelle opinioni dei tribunali e i tribunali generalmente danno un notevole rispetto alle linee guida dell’agenzia, ma i tribunali non sono vincolati da esse.

È difficile sopravvalutare l’importanza pratica delle agenzie. La validità o meno di una legge dipende in gran parte dall’agenzia esecutiva. L’EEOC inizialmente ha rifiutato di occuparsi della discriminazione di genere nonostante avesse il potere di farlo. In effetti, il Titolo VII non aveva alcun potere nemmeno contro la discriminazione razziale fino a quando non fu modificato nel 1972 per conferire all’EEOC il potere di agire (la Legge sulle Pari Opportunità di Lavoro del 1972).²³⁹ Le agenzie differiscono nel loro livello di indipendenza

politica; alcuni sono ospitati all’interno dell’esecutivo (come il Dipartimento per l’edilizia abitativa e lo sviluppo urbano e il Dipartimento del lavoro) mentre altri sono più indipendenti (come l’EEOC e la Federal Trade Commission (FTC)). Questi ultimi hanno poteri esecutivi oltre a poteri normativi. Possono condurre indagini e sporgere denuncia

in tribunale; alcuni hanno addirittura un proprio sistema giudiziario e talvolta vengono definiti de facto un quarto ramo del governo. In breve, sia l'interpretazione che l'applicazione degli statuti sono compiti condivisi dalle agenzie federali e dai tribunali. In genere lavorano insieme in armonia,¹⁰ ma la proliferazione di fonti giuridiche e di metodi di applicazione può portare a inefficienza e confusione.

Vale la pena menzionare altre due importanti fonti di politica: gli ordini esecutivi e le regole interne alle istituzioni. Gli ordini esecutivi sono direttive emanate dal presidente degli Stati Uniti. Originariamente inteso come un modo per gestire gli affari del governo, la vasta portata del governo federale ha fatto sì che esso diventi di fatto un potente strumento per l'attuazione delle politiche. Ad esempio, i precursori del titolo VII sotto forma di ordini esecutivi risalgono al 1941.

²⁴¹ Sebbene di portata molto più debole rispetto all'eventuale legislazione, essi illustrano la capacità dei presidenti di agire rapidamente mentre il Congresso potrebbe essere in fase di stallo.

Le istituzioni, siano esse pubbliche o private, possono stabilire regole o linee guida contro la discriminazione per i propri dipendenti che possono andare oltre i requisiti di legge. Ad esempio, chiedere a un candidato per un posto di lavoro il suo stato civile non è di per sé illegale.

Tuttavia, verrebbe interpretato come prova dell'intenzione di discriminare in una controversia legale.²⁴² Considerando questo (e il fatto che non c'è quasi mai un motivo legato al lavoro per tale indagine), molte organizzazioni vietano ai propri intervistatori di porre tali domande. Nel quotidiano, queste linee guida istituzionali sono le regole di non discriminazione più dirette a cui sono vincolati gli individui.

Caso di studio: l'evoluzione del Titolo IX

Il titolo IX della legge sugli emendamenti educativi del 1972 proibisce la discriminazione sessuale negli istituti scolastici che ricevono fondi federali. Nel 1975, il Dipartimento della sanità, dell'istruzione e del welfare (HEW) pubblicò i regolamenti finali che specificavano in dettaglio come sarebbe stato applicato il Titolo IX. Dal 1975, il governo federale ha pubblicato linee guida che chiariscono come interpretare e applicare tali regolamenti.

Due delle grandi domande che circondano il Titolo IX riguardavano cosa costituisce ricevere assistenza finanziaria federale e cosa costituisce discriminazione sessuale. Ognuno di questi aveva il potenziale per avere un impatto enorme sulla portata della legge. Nel 1984, la Corte Suprema stabilì che il Titolo IX era specifico per un programma: vale a dire che solo i programmi e le attività che ricevevano fondi federali diretti dovevano conformarsi. Ciò ha vanificato l'applicazione del Titolo IX: ad esempio, la maggior parte dei programmi sportivi non erano più coperti poiché non ricevevano direttamente fondi federali. In risposta, il Congresso elaborò un disegno di legge specificamente inteso a ribaltare questa decisione, ripristinando l'ampio campo di applicazione del Titolo IX, che conserva fino ad oggi.¹¹ [autore?][243]; (autore?) [244]]¹¹

¹⁰Tuttavia, dai tempi dell'amministrazione Trump la situazione è cambiata.²⁴⁰

¹¹Per inciso, l'istituzione al centro di questa controversia era il Grove City College, un college cristiano conservatore che non accettava direttamente alcun denaro federale nel tentativo di mantenere la propria autonomia. Dopo il 1988, sarebbe stato coperto dal Titolo IX perché i suoi studenti erano destinatari di prestiti e sovvenzioni federali, a dimostrazione dell'ampia portata dello Stato. Ad oggi, il college sfugge alle responsabilità del Titolo IX vietando ai suoi studenti di ricevere prestiti o sovvenzioni studentesche federali (come le sovvenzioni Pell).

Altre due importanti controversie riguardanti la portata del Titolo IX riguardano se le scuole siano responsabili delle molestie sessuali che avvengono nei campus e se sia vietata la discriminazione sulla base dell'orientamento sessuale e dell'identità di genere. A differenza della questione della copertura, questi continuano ad essere oggetto di un acceso dibattito legale e politico. Sul fronte delle molestie sessuali, la Corte Suprema ha affermato alla fine degli anni '90 che le scuole sono responsabili di creare un ambiente sicuro, compresa la prevenzione delle molestie da parte di altri studenti, ma "lo studente deve dimostrare che un funzionario della scuola con l'autorità di rispondere effettivamente sapeva ed è stato deliberatamente indifferente alle molestie". Le amministrazioni Obama, Trump e Biden hanno tutte introdotto linee guida o regolamenti su questa questione, ampliando e contraendo a loro volta la portata del Titolo IX.²⁴⁵ Un'altalena politica simile si è verificata rispetto alle protezioni LGBTQ, e l'ultima mossa è stata un'azione espansiva. interpretazione nel 2021 da parte del Dipartimento dell'Istruzione, rafforzata da una sentenza della Corte Suprema del 2020 che coinvolge anche la relazione tra discriminazione sessuale e orientamento sessuale, ma nel contesto della discriminazione sul lavoro.²⁴⁶

Come la legge concepisce la discriminazione

Esistono molti modi possibili per definire la discriminazione e tentare di raggiungere la non discriminazione. In questa sezione discuteremo di come la legge concepisce la discriminazione e di come cerca di bilanciare la non discriminazione con altri ideali.

Trattamento disparato e impatto disparato

Immaginate un datore di lavoro che rifiuti un candidato per un posto di lavoro e lo informi esplicitamente che questa decisione era dovuta a una caratteristica protetta. Un caso del genere sarebbe relativamente semplice da giudicare sulla base del testo dello statuto stesso ("Sarà una pratica di lavoro illegale per un datore di lavoro fallire o rifiutarsi di assumere un individuo... a causa della razza, del colore, della religione, sesso o origine nazionale.") Tuttavia, nella maggior parte dei casi di discriminazione, il comportamento del decisore è meno esplicito e le prove sono più circostanziate. Per affrontare questi problemi, i tribunali hanno creato due dottrine principali chiamate trattamento disparato e impatto disparato.

La disparità di trattamento si riferisce alla discriminazione intenzionale e corrisponde grosso modo alla concezione che la persona media ha del comportamento discriminatorio. Sussume il caso semplice descritto nel paragrafo precedente. Per i casi più circostanziali, la Corte Suprema ha istituito un cosiddetto quadro di trasferimento degli oneri ai sensi del Titolo VII (diritto del lavoro). In primo luogo, il ricorrente deve dimostrare un caso di discriminazione "prima facie" dimostrando di essere membro di una classe protetta, di essere qualificato per una posizione, gli è stata negata la posizione e la posizione è poi rimasta aperta o è stata assegnata a qualcuno non appartenente alla classe classe protetta. Se l'attore riesce in questo, il datore di lavoro deve fornire una ragione legittima e non discriminatoria per la decisione negativa. Spetta poi all'attore l'onere di provare che la motivazione addotta è mero pretesto di discriminazione.¹²

¹²Più precisamente, questo è solo uno degli ambiti possibili; lo descriviamo a scopo illustrativo.

La disparità di trattamento di solito comporta il ragionamento su quale azione il convenuto avrebbe intrapreso se le caratteristiche protette del querelante fossero state diverse, lasciando invariati tutti gli altri fatti del caso. Altrove in questo libro discutiamo perché, da un punto di vista tecnico, questi controfattuali di “inversione degli attributi” sono in contrasto con una comprensione sfumata della causalità e si traducono in fragili test di discriminazione. In ogni caso, l’importanza della causalità nei trattamenti disparati, in particolare il cosiddetto “ma-for-causation”, è aumentata a seguito di una decisione della Corte Suprema del 2020²³¹ che ha stabilito che è impossibile impegnarsi nella discriminazione basata sull’orientamento sessuale senza impegnarsi nella discriminazione sessuale immaginando un contesto controfattuale. in cui il sesso della vittima viene cambiato senza influenzare nient’altro, inclusa la preferenza di genere. Sebbene celebrato dal punto di vista dei diritti civili a causa delle sue implicazioni per i diritti LGBTQ, dovremmo tenere presente che ciò rappresenta una comprensione ristretta della causalità e la sua applicazione in altri scenari potrebbe portare a conclusioni non così favorevoli ai diritti civili.

A differenza del trattamento disparato, l’impatto disparato riguarda pratiche che hanno un effetto sproporzionato su una classe protetta, anche se non intenzionali. Ad alto livello, gli impatti disparati devono essere ingiustificati ed evitabili. Ciò viene nuovamente reso operativo attraverso un quadro di spostamento degli oneri. In primo luogo, il ricorrente deve dimostrare che esiste una differenza sproporzionata nei tassi di selezione tra i diversi gruppi. Se ciò può essere dimostrato, il datore di lavoro ha la possibilità di spiegare se il motivo delle diverse percentuali di selezione ha una giustificazione aziendale. L’onere ricade quindi sul ricorrente per dimostrare che esiste una “pratica lavorativa alternativa” che avrebbe raggiunto gli obiettivi del datore di lavoro pur essendo meno discriminatoria.

Un modo di pensare agli impatti disparati è come un modo per “annusare” la discriminazione intenzionale ben nascosta concentrandosi sui suoi impatti, che sono più facilmente osservabili. In effetti, il caso che ha portato alla dottrina riguardava un datore di lavoro che ha introdotto test attitudinali per la promozione proprio il giorno in cui è entrato in vigore il Civil Rights Act del 1964, che proibiva la discriminazione sul lavoro basata sulla

razza.²⁴⁸ Ma si ritiene anche che vi siano impatti disparati motivati da una considerazione di giustizia distributiva, cioè minimizzando le disuguaglianze ingiustificate nei risultati. In questo senso, l’impatto disparato corrisponde grosso modo alla visione mediana delle pari opportunità di cui abbiamo discusso nel capitolo 4. L’impatto disparato cerca di costringere i decisori a trattare persone apparentemente dissimili in modo simile nella convinzione che la loro attuale dissomiglianza sia il risultato di ingiustizie passate.. Mira a compensare almeno alcuni degli svantaggi subiti per ragioni ingiuste. Infatti, nel caso sopra menzionato, la Corte Suprema ha sottolineato che la disparità razziale nei risultati dei test attitudinali potrebbe essere spiegata dalle disuguaglianze nel sistema educativo. Ma la dottrina dell’impatto disparato si è evoluta nel corso degli anni e si ritiene che la misura in cui riflette la giustizia distributiva, invece di uno strumento per illuminare una discriminazione ben nascosta, sia diminuita nel tempo.

Mentre abbiamo discusso queste due dottrine nel contesto dell’occupazione

Il quadro completo – incluso il gioco su quale quadro scegliere – è un vasto pantano.²⁴⁷

diritto, si trovano in ciascuno dei sei ambiti discussi nella prima sezione.

L'impatto disparato è stato così centrale nella comprensione giuridica della discriminazione che è stato successivamente incorporato negli statuti, in particolare nell'ADA (legge sulla disabilità), ma anche nel Titolo VII (legge sulla parità di lavoro) attraverso un emendamento del 1991. Ma la Corte Suprema non ha esteso la dottrina a situazioni in cui le leggi o le procedure dello Stato (piuttosto che di attori privati) violano la clausola di pari protezione se hanno uno scopo discriminatorio. In altre parole, non esiste un equivalente della dottrina dell'impatto disparato per gli attori statali, ma solo un trattamento disparato.

Un'importante osservazione generale sulla legge antidiscriminatoria – soprattutto per i lettori che potrebbero essere abituati a pensare all'equità in termini di proprietà statistiche dei risultati dei processi decisionali – è che la legge si occupa principalmente dei processi stessi. Allo stesso modo, anche il modo in cui i tribunali valutano le prove è altamente procedurale, al punto che può sembrare tangenziale alla questione sostanziale se abbia avuto luogo o meno una discriminazione.²⁴⁹ Anche l'impatto più disparato, nonostante sia motivato in parte da nozioni distributive di giustizia, è trattati in modo formale e procedurale. A dimostrazione della centralità dell'elemento procedurale, il manuale legale del Dipartimento di Giustizia per dimostrare le affermazioni di diverso impatto del Titolo VI è lungo oltre 20.000 parole.²⁵⁰ Ci sono molte possibili ragioni per cui la legge si concentra sul processo. Il primo è storico: gli statuti rispondevano principalmente a discriminazioni schiette e negazioni formali di

opportunità in contrapposizione a fenomeni statistici più subdoli. Si adatta meglio anche al funzionamento della legge: la definizione di discriminazione non può essere separata dalla procedura per dimostrarla in tribunale. Una terza ragione è politica: è più facile raggiungere il consenso su processi equi che su quale sia il giusto risultato distributivo. Infine, a livello pragmatico, riflette l'attenzione della legge alle sfumature dei luoghi di lavoro e di altre istituzioni, rispetto alle quali i criteri di equità statistica sembrano rozzi ed eccessivamente semplificati.

Evitare oneri eccessivi per i decisor

Un tema ricorrente è quanto onere gravante sui decisor sia giustificato nel perseguimento di obiettivi di equità. Ad esempio, la realizzazione di alloggi per i dipendenti disabili comporta dei costi per un'azienda.

In generale, la legge dà sostanziale rispetto agli interessi del decisore. Ciò è stato più volte chiarito dai legislatori e dai tribunali in vari momenti. Ad esempio, la Commissione Giustizia della Camera ha affermato in merito al ruolo dell'EEOC al momento della creazione dell'agenzia: "le prerogative del management e le libertà sindacali devono essere lasciate indisturbate nella massima misura possibile."²⁵¹ La Corte Suprema ha chiarito nel 2015 che "Il FHA (Fair Housing Act) non è uno strumento per costringere le autorità competenti in materia di edilizia abitativa a riordinare le loro priorità."²⁵²

Un'eccezione è l'Americans with Disabilities Act, che impone sostanziali requisiti di conformità a un ampio numero di aziende e governi. Ciò non dovrebbe sorprendere dal momento che la legge ha cercato di creare cambiamenti strutturali nella società, in particolare nell'ambiente edificato. L'ADA ha una difesa "indebita difficoltà" rispetto al requisito secondo cui i datori di lavoro forniscono "accomodamenti ragionevoli" ai lavoratori qualificati

dipendenti con disabilità, ma i tribunali sembrano tollerare un onere più elevato rispetto , ad esempio, al Titolo VII.¹³ A titolo illustrativo, i dipendenti ciechi hanno citato in giudizio i loro datori di lavoro per non aver fornito lettori pagati per quattro ore della giornata lavorativa; la corte si è schierata dalla parte dei ricorrenti, lasciando intendere che un aumento di circa il 50% nel costo dei dipendenti per il datore di lavoro non costituiva un indebito disagio.²⁵³ Ci affrettiamo ad aggiungere che un indebito disagio è un test multifattoriale e non vi è alcuna chiara o soglia di costo uniforme; il costo è raramente il fattore determinante.

Esistono diverse potenziali giustificazioni degli oneri che gravano sui decisorii nel mondo degli studi accademici e nella storia legislativa. Spesso le responsabilità dei decisorii sono giustificate facendo appello ai diritti umani di coloro che vengono danneggiati, piuttosto che da un'analisi economica. Ad esempio, l'ADA ha cercato di intervenire nella discriminazione contro le persone disabili che spesso incide sui loro mezzi di sussistenza e talvolta costa la vita.²⁵³ In alternativa, gli oneri sono talvolta giustificati perché rappresentano solo un costo "de minimis". Ad esempio, il Titolo VII impone ai datori di lavoro di adeguarsi alle convinzioni religiose dei dipendenti, ma non se ciò comporta costi superiori a quelli minimi.

Tra questi due tipi di casi ce ne sono altri in cui è necessario un più attento equilibrio tra benefici e costi.¹⁴ Diamo qui alcuni brevi esempi.

- Esteriorità positiva. Una speranza dietro l'Americans with Disabilities Act era che avrebbe reso più facile per le persone disabili entrare nel mondo del lavoro e contribuire all'economia generale.¹⁵ •

La regolamentazione come azione collettiva. Il Titolo II del Civil Rights Act del 1964 proibisce la discriminazione nei luoghi di alloggio pubblico (ad esempio, ristoranti). Uno dei motivi principali per cui tali strutture discriminano le minoranze è dovuto ai pregiudizi dei loro clienti bianchi. Il Titolo II ha consentito loro di smettere di discriminare , di ottenere affari da clienti minoritari senza incorrere in perdite di affari da parte dei loro clienti bianchi; pertanto, la legge non ha imposto loro un onere ma ha piuttosto creato un'opportunità.²⁵⁴ Allo stesso modo, consideriamo l'assicurazione. In assenza di regolamentazione, se un assicuratore evita di calibrare i premi in base al rischio nell'interesse dell'equità, potrebbe fallire. Ma se tutte le imprese sul mercato vincolassero il loro comportamento allo stesso modo alle leggi antidiscriminatorie, non potrebbero più affermare di trovarsi in una situazione di svantaggio competitivo.

- Evita i costi più economici. Il principio dell'evitatore del costo più basso o dell'evitatore del costo minimo attribuisce la responsabilità del danno alla parte che può evitare il danno al costo più basso. È il motivo per cui le aziende, in una certa misura, si assumono la responsabilità per discriminazioni o molestie commesse dai propri dipendenti. Se un datore di lavoro

¹³ Inizialmente l'ADA richiedeva un limite più elevato: un accomodamento avrebbe dovuto mettere a repentaglio la continuazione dell'attività del datore di lavoro.²⁵³ In effetti, la parte della legge che si applica ai servizi pubblici prevede un limite più elevato per il convenuto: l'accomodamento deve fondamentalmente alterare la natura del servizio o del programma.

¹⁴Questi tipi di domande vengono studiate nel campo del diritto e dell'economia, che applica la microeconomia per spiegare gli effetti delle leggi.

¹⁵In effetti, la quota di disabili occupati è diminuita dall'entrata in vigore della I ADA, anche se l'effetto causale è tutt'altro che chiaro.

è costretto a internalizzare i costi delle molestie discriminatorie commesse dai suoi dipendenti, secondo la teoria economica standard, investirà in precauzioni fino al punto in cui queste non saranno più giustificate in termini di costi.²⁵⁵

- Correggere l'irrazionalità. Alcuni commentatori suggeriscono che la dilagante discriminazione contro le donne prima dell'approvazione dell'ECOA fosse un comportamento irrazionale da parte dei creditori, e che le donne costituissero in realtà un buon rischio di credito . sui creditori.

Limiti della legge nel contrasto alla discriminazione

Quanto è stata efficace la legge antidiscriminazione degli Stati Uniti? Lo scenario migliore è che la possibilità di sanzioni abbia sufficientemente scoraggiato i potenziali discriminatori da far crollare i tassi di discriminazione e, nei pochi casi di discriminazione rimasti , le vittime riescano a ottenere risarcimento attraverso i tribunali. Lo scenario peggiore è che le leggi non abbiano praticamente avuto alcun effetto e qualsiasi riduzione delle disparità dopo la loro approvazione può essere attribuita ad altri fattori, come ad esempio il minor successo dei discriminatori sul mercato.

La realtà è da qualche parte nel mezzo. Valutare rigorosamente l'effetto delle leggi è un problema controfattuale complicato ed è soggetto a molte incertezze e dibattiti. Tuttavia, ci sono molte prove che suggeriscono un effetto positivo. Ad esempio, uno studio ha utilizzato un esperimento naturale per valutare l'impatto del Titolo VII sulle opportunità di lavoro per gli afroamericani rispetto ai bianchi americani. Ha dimostrato che l'occupazione relativa degli afroamericani è aumentata maggiormente nelle industrie e nelle regioni con una percentuale maggiore di imprese che erano state recentemente coperte dal Titolo VII dell'Equal Employment Opportunity Act²⁵⁷⁻²⁵⁸ del 1972.

Anche se i vantaggi non sono stati trascurabili, l'efficacia della legge antidiscriminazione è ridotta per molte ragioni di cui ora discutiamo. Ciò motiva la nostra visione secondo cui il lavoro sull'equità algoritmica non dovrebbe considerare l'approccio adottato nella legge antidiscriminatoria come un dato di fatto, ma dovrebbe invece riconnettersi con i fondamenti morali dell'equità.

Gli oneri sulle vittime di discriminazione

La legge impone una serie di oneri alle vittime di discriminazione nel caso in cui desiderino ricorrere alle vie legali. Utilizzeremo il mercato del lavoro come esempio, ma le nostre osservazioni si applicano anche ad altri contesti.

Per cominciare, l'intervento legale viene avviato dalla vittima, non dal governo,¹⁶ e non può iniziare finché le vittime non hanno già subito la discriminazione. I regolatori non esaminano in modo prospettico le pratiche occupazionali, a differenza di altri settori

¹⁶Ci sono limitate eccezioni a questo principio generale, come la capacità delle agenzie di contrasto di portare avanti casi di "modello o pratica" contro i discriminatori recidivi. Un esempio importante è l' accordo del Dipartimento di Giustizia contro il Dipartimento di Polizia della Pennsylvania. Stati Uniti contro Pennsylvania e Polizia di stato della Pennsylvania, n.1:14-cv-01474-SHR (MD Pa. 29 luglio 2014), <https://www.justice.gov/sites/default/files/crt/legacy/2014/07/31/pennsylvaniapdcomp.pdf>.

di legge come la regolamentazione farmaceutica in cui i farmaci devono essere accuratamente testati prima di essere ammessi sul mercato. Inoltre, esiste una fondamentale asimmetria informativa tra aziende e dipendenti (o candidati al lavoro). Le vittime potrebbero anche non essere consapevoli di aver subito discriminazioni. Dopo tutto, i candidati e i dipendenti non hanno visibilità diretta sul processo decisionale dei datori di lavoro e le aziende non hanno bisogno di fornire una giustificazione per una decisione sfavorevole di assunzione o promozione. Solo nel settore del credito esiste una qualche forma di obbligo di trasparenza.¹⁷ Anche se una

vittima viene a conoscenza della discriminazione, deve affrontare ostacoli che potrebbero dissuaderla dal fare causa. Il contenzioso comporta ulteriore angoscia mentale. Le vittime possono anche essere scoraggiate dagli elevati costi finanziari del contenzioso.¹⁸ Le cause legali normalmente impiegano diversi anni per giungere a una conclusione, momento in cui la carriera della vittima può subire una battuta d'arresto significativa e irreparabile. Se la vittima rimane presso l'azienda dopo aver intentato una causa, nel migliore dei casi dovrà affrontare una situazione scomoda e potenziali ritorsioni da parte del datore di lavoro (anche se le leggi vietano specificamente le ritorsioni, rimane un risultato comune di azioni legali per discriminazione). E se la vittima cerca lavoro altrove, i futuri datori di lavoro potrebbero considerare negativamente il fatto che il candidato abbia citato in giudizio un precedente datore di lavoro per discriminazione.

Le vittime che decidono di fare causa devono affrontare una serie di ostacoli procedurali. Se il datore di lavoro dispone di procedure interne di reclamo, alla vittima potrebbe essere richiesto di provarle prima di fare causa (o rischiare di perdere le sue pretese). Un altro prerequisito per intentare una causa è presentare un reclamo amministrativo alla Commissione per le pari opportunità di lavoro subito dopo l'inizio della discriminazione. Il requisito della tempestività spesso pone le vittime in un doppio vincolo a causa della necessità di esaurire i canali interni. Inoltre, rende difficile raccogliere le prove necessarie per prevalere in tribunale.²⁵⁹

260

Ciò ci porta all'ultima e più grave difficoltà che i querelanti devono affrontare, ovvero l'onere della prova. A dire il vero, lo standard di prova che il querelante deve soddisfare nei casi di discriminazione è la "preponderanza delle prove", che significa più probabile che no, che è inferiore allo standard nei casi penali. Ma anche questo standard si è rivelato scoraggiante. Secondo Katie Eyer, "la legge antidiscriminazione è un'area tecnica molto rigida della legge, in cui qualsiasi miriade di dottrine tecniche può portare al licenziamento. I tribunali affrontano la questione della discriminazione come se fosse un complesso puzzle legale, in cui qualsiasi pezzo fuori posto deve comportare il rigetto delle pretese dei querelanti."²⁴⁹

Nello specifico, in casi di trattamento disparato, i tribunali hanno creato numerose dottrine favorevoli all'imputato. Secondo la dottrina delle "osservazioni vaganti", i commenti discriminatori fatti dal datore di lavoro nei confronti del querelante non costituiscono prova di intento discriminatorio a meno che non vi sia un nesso causale sufficientemente chiaro con la decisione stessa. Secondo la difesa dello "stesso attore", se il datore di lavoro era disposto ad assumere l'attore in precedenza, si considera prova che il datore di lavoro non ha alcuna responsabilità

¹⁷Tuttavia, il passaggio a strumenti algoritmici nelle assunzioni ha aperto la possibilità di requisiti di trasparenza e revisione ex ante. Una coalizione di organizzazioni per i diritti civili ha sostenuto tali pratiche in un documento che definisce una serie di principi sui diritti civili per l'assunzione di tecnologie di valutazione.²⁵⁸ Discuteremo idee emergenti come le valutazioni di impatto algoritmico nella sezione finale di questo capitolo.

¹⁸Ciò può essere mitigato in alcuni casi se gli studi legali sono disposti a essere pagati sulla base di una commissione di contingenza – in cui vengono pagati solo in caso di risultato favorevole come percentuale fissa dei danni recuperati – o se lo statuto prevede una compensazione collettiva o azioni di classe.

intento discriminatorio nei confronti del ricorrente. Secondo la regola della “credenza onesta”, un caso può essere archiviato sommariamente se il datore di lavoro “crede onestamente” nelle ragioni della decisione, anche se in seguito si può dimostrare che sono “errate, insensate, banali o infondate”.

Con effetti disparati, una serie di fattori sovrapposti viene schierata contro il querelante. Anche se non è necessario stabilire l'intento, esiste una nuova serie di requisiti: identificare una politica o una pratica specifica che ha causato la decisione sfavorevole sull'assunzione; compilare le statistiche necessarie per dimostrare che la politica ha un impatto dispari; e confutare la difesa del datore di lavoro secondo cui la politica è giustificata dal contesto lavorativo. Il terzo punto rappresenta un ostacolo particolarmente grave per i ricorrenti in quanto sono strutturalmente scarsamente posizionati per identificare una pratica lavorativa alternativa, poiché non hanno la conoscenza degli aspetti interni dell'attività svolta dal

datore di lavoro.^{261, 253} Il risultato netto di questi ostacoli per i ricorrenti è che le loro probabilità di successo al processo sono estremamente basse. Katie Eyer riassume i dati del progetto Uncertain Justice: “su 100 querelanti per discriminazione che portano avanti le loro richieste fino alla conclusione (vale a dire, non risolvono o respingono volontariamente le loro richieste), solo 4 ottengono una qualche forma (de minimis o meno) di sollievo. . . . Queste probabilità possono essere propriamente caratterizzate come scandalosamente scarse e si estendono (con piccole differenze)

a ogni categoria di attori discriminatori, inclusi razza, sesso, età e disabilità. troppo facile da presentare e troppo favorevole ai querelanti, una posizione che respingiamo. Selmi esamina criticamente questa percezione e rileva che è prevalente tra i giudici; correggere questo squilibrio percepito può infatti essere una delle ragioni per la creazione di numerosi ostacoli per i querelanti.²⁶² Che si sottoscriva o meno l'idea secondo cui molte “cause per molestie” vengono intentate da querelanti che accusano di discriminazione, è vero che i tribunali sono molto tesi e i giudici sono diffidenti nei confronti delle decisioni che potrebbero aprire le “porte” a cause legali. Ciò suggerisce che è improbabile che gli oneri di cui abbiamo discusso in precedenza scompaiano.^{263, 260}

La difficoltà di riforme sostanziali e strutturali attraverso interventi procedurali

Anche se il rispetto della legge antidiscriminatoria è elevato e i rimedi giuridici sono facilmente ottenibili, potrebbero esserci limiti ancora più fondamentali all'efficacia della legge. In che misura i limiti formali che la legge impone agli individui e alle organizzazioni conducono a una società giusta? Quanto è grande il divario tra le nozioni legali e morali di ingiustizia?

Stephen Halpern inquadra la questione così:²⁶⁴

Nel tradurre un problema sociale nel “linguaggio” della legge, i giuristi devono inquadrare la loro analisi in termini di concetti, questioni, domande e rimedi inventati che il sistema legale riconosce e considera legittimi . In quella traduzione, come in ogni traduzione, ci sono costrizioni e distorsioni. Inquadrare un problema sociale come una questione giuridica produce una trasformazione della questione stessa – una riconcettualizzazione del problema, producendo domande e preoccupazioni uniche che diventano per prime il fulcro dell'attenzione.

il dibattito giuridico e successivamente tendono a dominare la discussione pubblica. Quando i problemi razziali vengono riformulati come questioni di diritti legali, il dialogo che ne risulta non coglie la complessità e la sottigliezza di tali problemi né consente di prendere in considerazione la gamma più completa di rimedi ad essi. Inevitabilmente, le esigenze e i limiti del processo legale alterano il discorso pubblico e la comprensione delle questioni razziali vitali.

Il libro di Halpern riguarda la disuguaglianza razziale nell'istruzione; l'esempio principale della sua tesi è lo sforzo compiuto per porre fine alla segregazione delle scuole pubbliche senza prestare molta attenzione alla qualità dell'istruzione ricevuta dagli studenti neri nelle scuole integrate. Allo stesso modo, il contenzioso del Titolo VI si è concentrato sulle procedure per l'elaborazione delle denunce di discriminazione presentate al governo federale, piuttosto che sui meccanismi per rivendicare diritti sostanziali a un'istruzione di qualità comparabile. Egli sostiene che “[f]ew, se non nessuno, dei fattori che hanno un impatto sul rendimento scolastico sono regolati da” diritti legali “o sono facilmente traducibili in una questione di” discriminazione razziale “”. Fornisce due ragioni per cui le disuguaglianze persistono nonostante i rimedi formali della legge: la segregazione di fatto delle città americane e il fatto che le differenze accademiche spesso derivano dall'instabilità domestica e da altre disparità sociali, economiche e sanitarie.

Sebbene l'effetto della desegregazione scolastica negli Stati Uniti sia un argomento vasto, il punto più ampio è che le limitazioni del processo legale limitano ciò che è realizzabile e modellano anche la nostra comprensione dei problemi stessi. Un altro esempio di ciò viene dal libro *Color of Law* di Richard Rothstein:²²²

Sebbene la maggior parte degli afroamericani abbia sofferto a causa di [politiche abitative governative storicamente razziste], non riesce a identificare, con la specificità che un caso giudiziario richiede, il momento particolare in cui sono stati vittime. Ad esempio, molti veterani afroamericani della seconda guerra mondiale non fecero domanda per mutui garantiti dal governo per gli acquisti suburbani perché sapevano che l'amministrazione dei veterani li avrebbe respinti a causa della loro razza, quindi le domande erano inutili.

Quei veterani quindi non guadagnavano ricchezza dall'apprezzamento del patrimonio immobiliare come facevano i veterani bianchi, e i loro discendenti non potevano quindi ereditare quella ricchezza come facevano i discendenti dei veterani bianchi. Con meno ricchezza ereditata, gli afroamericani oggi sono generalmente meno capaci dei loro coetanei bianchi di permettersi di frequentare buone università. Se uno di quei discendenti afroamericani venisse a sapere che il motivo per cui i suoi nonni sono stati costretti ad affittare appartamenti in aree urbane sovraffollate era che il governo federale proibiva incostituzionalmente e illegalmente alle banche di concedere prestiti agli afroamericani, il nipote non avrebbe il diritto di presentare domanda una causa; né lui o lei sarebbe in grado di nominare una determinata parte da casa, i danni potrebbero essere recuperati.

Un altro impulso verso il proceduralismo deriva dall'interazione del sistema giudiziario con le procedure interne delle organizzazioni. Secondo la teoria dell'endogeneità legale, le organizzazioni attuano protezioni procedurali, come la formazione sulla diversità

programmi, con il presunto scopo di frenare la discriminazione; col passare del tempo, i tribunali arrivano gradualmente a confondere queste attività orientate al rispetto procedurale e simbolico per misure sostanziali; ma una volta che queste stesse misure simboliche acquisiscono significato legale, le preoccupazioni sostanziali sono state spinte fuori dall'ambito del legittimo

dibattito.²⁶⁵ Inoltre, anche un cambiamento sostanziale a livello delle singole organizzazioni potrebbe non implicare un cambiamento strutturale, cioè un cambiamento dei fattori sottostanti nella società che producono innanzitutto le disparità. Anche se un datore di lavoro raggiungesse la parità statistica nei tassi di assunzione e promozione, i tassi di candidatura potrebbero essi stessi riflettere disuguaglianze di opportunità nella società e/o discriminazioni a livelli o fasi precedenti del sistema, e c'è poco che la legge può fare per obbligare i singoli decisori a porre rimedio a queste disuguaglianze.

Gli interventi giuridici i cui effetti sono sia sostanziali che strutturali sono rari. Un esempio notevole è l'impatto del titolo IX sull'atletica femminile. La legge è stata interpretata non solo in modo da vietare la discriminazione in senso stretto, ma anche in modo da richiedere equità in una serie di settori quali borse di studio, coaching e strutture. Probabilmente come risultato di questi interventi, il prestigio dell'atletica femminile negli Stati Uniti è gradualmente aumentato, indebolendo la gerarchia di genere nell'atletica, portando a una maggiore parità nell'atletica anche al di fuori del contesto collegiale.¹⁹ In generale, tuttavia, questi tipi di interventi sostanziali si sono finora rivelati meno realizzabili di quelli formali, in parte a causa dei finanziamenti che richiedono.

Sebbene abbiamo contrapposto gli interventi procedurali agli interventi sostanziali e strutturali di cui sopra, il confine tra loro può essere confuso, e i primi possono almeno funzionare come un punto di appoggio per i secondi. Nella misura in cui la disuguaglianza persiste a causa di politiche radicate che mantengono una distribuzione ineguale delle risorse, gli interventi procedurali che consentono ai membri di gruppi storicamente oppressi di raggiungere posizioni di autorità potrebbero consentire loro di alterare in modo più efficace queste politiche. Gli interventi procedurali possono contribuire a ridurre la capacità dei gruppi già avvantaggiati di usurpare il pieno controllo sul processo di elaborazione delle politiche. Tuttavia, questo è ben lungi dall'essere un percorso ideale per il cambiamento, poiché pone l'onere di promuovere gli interessi di gruppi specifici sugli individui che appartengono a quei gruppi.

Un altro contrasto apparente è tra la legge sulla discriminazione e le politiche redistributive, vale a dire che il governo tassa direttamente alcuni attori e ridistribuisce tali fondi al gruppo svantaggiato. Ma la legge sulla discriminazione può essere intesa come un meccanismo che scarica in una certa misura l'onere economico della rettifica delle ingiustizie passate sui datori di lavoro, sui finanziatori, ecc. In un certo senso, questo potrebbe essere simile a una politica di tassazione dei datori di lavoro e utilizzo di tali fondi per sostenere i gruppi che in passato hanno subito discriminazioni.

Le politiche di azione affermativa, in particolare, occupano uno spazio che è esattamente a metà tra la non discriminazione formale e le politiche redistributive. Un esempio di tale politica potrebbe essere un programma di formazione professionale offerto da un datore di lavoro che favorisca i gruppi con minore accesso alle opportunità.²⁶⁶ Tuttavia, tranne rari casi, la legge non impone azioni positive da parte di soggetti privati ma si limita a consentirle.

¹⁹Ciò non vuol dire che l'equità sia stata raggiunta nell'atletica femminile o nei programmi atletici generale: gli orribili scandali degli abusi sessuali ci ricordano che c'è ancora molta strada da fare.

Più comunemente visti sono i requisiti affermativi per i governi. Il Fair Housing Act, oltre ai mandati di non discriminazione, richiede all'HUD e ai destinatari dei fondi federali da parte dell'HUD di "promuovere in modo affermativo" le politiche e gli scopi della legge. Ciò potrebbe consentire, ad esempio, alloggi sovvenzionati nelle comunità ad alto reddito che aprono l'accesso a scuole e servizi di qualità superiore.

Tuttavia, questa parte della FHA è rimasta in gran parte inattiva. Pertanto, almeno una parte dei limiti della legge nel creare cambiamenti significativi può essere attribuita alla mancanza di volontà politica di agire pienamente in base alle leggi esistenti, piuttosto che a una limitazione intrinseca del sistema legale.

Regolamentare l'apprendimento automatico

Sebbene la legge antidiscriminazione statunitense sia antecedente all'uso diffuso dell'apprendimento automatico , è altrettanto applicabile se un decisore utilizza l'apprendimento automatico o altre tecniche statistiche. Detto questo, l'apprendimento automatico introduce molte complicazioni nell'applicazione di queste leggi. Queste complicazioni sono oggetto di un acceso dibattito negli studiosi giuridici e molti studiosi temono che la legge esistente possa essere inadeguata per affrontare i tipi di discriminazione che sorgono quando è coinvolto l'apprendimento automatico. Allo stesso tempo, esiste anche l'opportunità di esercitare nuovi strumenti normativi per frenare la discriminazione algoritmica. C'è poca giurisprudenza su questo argomento, quindi la nostra discussione di questi problemi si baserà sulla dottrina giuridica. Come in precedenza, la nostra discussione è incentrata sugli Stati Uniti, ma in alcuni punti tocchiamo il regolamento generale sulla protezione dei dati (GDPR) dell'UE .

Trattamento disparato

Ricordiamo che le due principali dottrine antidiscriminatorie sono il trattamento disparato e l'impatto disparato. La disparità di trattamento riguarda principalmente l' intento esplicito di discriminare sulla base di caratteristiche giuridicamente tutelate; al contrario, l'impatto disparato si concentra sul processo decisionale in cui non vi è alcun intento esplicito di discriminare, ma dove anche le decisioni prese sulla base di caratteristiche apparentemente benigne si traducono tuttavia in disparità ingiustificate rispetto a caratteristiche legalmente protette.

La maggior parte delle segnalazioni di discriminazione nell'apprendimento automatico riguardano casi di discriminazione non intenzionale piuttosto che intenzionale. Inoltre, è improbabile che gli sviluppatori di sistemi di apprendimento automatico che intendano discriminare facciano affidamento esplicitamente su attributi protetti a causa della facile disponibilità di proxy. Quando ciò accade, può essere difficile dimostrare che vi fosse l'intento di mascherare la discriminazione. Per questi motivi, raramente si invoca un trattamento disparato e l'impatto disparato è considerato molto più rilevante. Torneremo presto all'impatto più disparato. Ma una questione importante che implica un trattamento disparato riguarda i sistemi che si basano esplicitamente sull'attributo protetto per correggere le distorsioni dei dati o mitigare gli effetti della discriminazione passata. Ciò costituisce un trattamento disparato? In altre parole, la legge impone limiti agli interventi di equità algoritmica?

La risposta è sfumata. Un aspetto relativamente positivo della legge è che le quote di selezione sono incostituzionali. In termini di machine learning, ciò corrisponde approssimativamente alla differenza tra le tecniche che mirano a imporre la parità e quelle che si limitano a penalizzare la disparità durante la fase di ottimizzazione. Quest'ultimo tipo di tecnica è analogo a un processo che è consapevole della razza e valorizza la diversità, ma consente comunque che la distribuzione finale vari a seconda dell'insieme dei candidati. È utile, come sempre, ricordare che le distinzioni tecniche raramente si associano chiaramente a determinazioni legali.

Esiste anche una grande differenza tra una decisione individuale presa sulla base di una caratteristica protetta e una politica complessiva che tenga conto degli interessi dei gruppi protetti. La disparità di trattamento si applica principalmente al primo tipo di decisione. Ciò è simile alla distinzione tra l'uso di un attributo protetto durante la formazione rispetto al momento della prova, sebbene, ancora una volta, questa distinzione di per sé sia lungi dall'essere giuridicamente determinante.

Un caso della Corte Suprema che viene spesso citato come esempio della tensione tra trattamento disparato e impatto disparato (e le insidie del processo decisionale consapevole della razza) è Ricci v. DeStefano. Il caso è sorto perché i vigili del fuoco di New Haven hanno annullato un esame promozionale dopo aver scoperto che i vigili del fuoco neri avevano un tasso di passaggio inferiore rispetto ai vigili del fuoco bianchi. Il dipartimento temeva che si sarebbe aperto a responsabilità di impatto disparate. Ma è stato poi denunciato dai vigili del fuoco bianchi e ispanici che avrebbero ottenuto la promozione in base all'esame. La corte ha concordato con i ricorrenti che il dipartimento aveva adottato un trattamento disparato nei loro confronti.

Pauline Kim sottolinea una caratteristica distintiva cruciale del caso Ricci: i querelanti avevano già investito molto tempo e denaro nello studio per l'esame, e quindi le azioni del dipartimento hanno provocato un danno concreto a individui specifici.

Dal punto di vista di Kim, la logica della Corte non si applicherebbe quando un datore di lavoro apporta una modifica alle sue pratiche di assunzione al fine di evitare il potenziale di impatti

disparati.²⁶⁷ Infine, anche se una pratica costituisce prima facie un trattamento disparato, potrebbe essere legale se fa parte di un valido programma di azione affermativa, cioè che mira a porre rimedio alle discriminazioni del passato. Nell'occupazione, la Corte Suprema ha stabilito che i programmi di azione affermativa basati sulla razza o sul genere sono validi se cercano di eliminare gli "squilibri manifesti" in "categorie lavorative tradizionalmente segregate" e non "intralciano inutilmente" gli interessi di altri candidati. Alcuni studiosi hanno sostenuto che ciò dovrebbe valere anche per l'azione affermativa algoritmica volontaria.²⁶⁸

Impatto disparato

Per comprendere in che modo l'impatto disparato si applica al processo decisionale statistico, dobbiamo analizzare la dottrina giuridica. Il quadro di trasferimento degli oneri stabilito dalla Corte Suprema per le denunce di discriminazione sul lavoro del Titolo VII funziona come segue.²⁶⁹ In primo luogo, il ricorrente deve stabilire un caso prima facie mostrando una differenza sufficiente nei tassi di selezione tra i diversi gruppi. Ciò che costituisce a

differenza sufficiente non è chiara. L'EEOC ha stabilito come linea guida una soglia di quattro quinti (ovvero una differenza del 20%), ma questa non è una regola rigida. In un mondo basato sui big data, alcuni commentatori hanno sostenuto che il criterio dovrebbe basarsi sulla significatività statistica della differenza piuttosto che sulla grandezza.²⁰

Se il querelante riesce a dimostrare una differenza sufficiente, l'onere passa al convenuto, che deve quindi dimostrare che la pratica contestata è “legata al lavoro” e coerente con la “necessità aziendale”. Se l'imputato può dimostrarlo, allora il querelante può ancora vincere dimostrando che esiste una “pratica di lavoro alternativa” che avrebbe raggiunto gli obiettivi del datore di lavoro pur essendo meno discriminatoria.

Il passaggio critico dal punto di vista del processo decisionale statistico è la questione della necessità aziendale. Un modo in cui il datore di lavoro può dimostrarlo è attraverso “dati empirici che dimostrano che la procedura di selezione è predittiva o significativamente correlata con elementi importanti della prestazione lavorativa”. Poiché l'apprendimento automatico è una tecnica per stabilire la validità predittiva, commentatori come Baracas e Selbst suggeriscono che ciò rappresenta un livello estremamente basso per i datori di lavoro.⁹ Finché la variabile target utilizzata in un modello predittivo è presumibilmente correlata al lavoro, il requisito è soddisfatto .

D'altro canto, Pauline Kim sostiene che il Titolo VII può effettivamente affrontare gli effetti discriminatori dell'apprendimento automatico, sulla base di una lettura attenta dello statuto.²⁷⁰ Tuttavia, la dottrina che si è sviluppata dopo la sua approvazione è inadeguata ad affrontare il problema apprendimento. Ad esempio, l'obbligo per il querelante di identificare una pratica lavorativa specifica che ha causato la disparità si è sviluppato in un'epoca in cui le prove scritte erano il veicolo principale per impatti disparati. Ma quando è in gioco un modello statistico, soprattutto se non interpretabile e che utilizza un gran numero di caratteristiche, non è chiaro cosa il querelante dovrebbe identificare. Pertanto, la dottrina dovrà evolversi se il Titolo VII vuole affrontare l'apprendimento automatico discriminatorio.

Un altro problema specifico del processo decisionale automatizzato deriva dal fatto che il software di solito non è sviluppato internamente dal decisore ma piuttosto da aziende esterne specializzate. Ad esempio, aziende come Hirevue e Pymetrics offrono strumenti per automatizzare parte del processo di assunzione e Upstart fornisce un modello predittivo per la sottoscrizione dei prestiti. In questi casi, chi dovrebbe assumersi la responsabilità?

Nel diritto del lavoro, sono i datori di lavoro, non i venditori, ad essere legalmente responsabili.²⁷¹ Ma i datori di lavoro (e altri clienti di questi strumenti) si oppongono a questa affermazione poiché di solito non hanno le competenze per condurre una validazione statistica. Trasferire parte o tutta la responsabilità dai clienti ai venditori avrebbe pro e contro dal punto di vista antidiscriminatorio. Ciò potrebbe significare che i fornitori diventeranno molto più attenti nel testare le loro offerte. D'altro canto, anche se uno strumento è stato ampiamente testato per un impatto disparato, potrebbe funzionare in modo diverso nel contesto di un particolare bacino di candidati di un particolare datore di lavoro. Inoltre, i ricorrenti potrebbero incontrare ancora più difficoltà nel dimostrare una pratica lavorativa alternativa.

Mentre la disparità di trattamento e l'impatto disparato sono i due pilastri principali della legge antidiscriminatoria, quando si tratta di decisioni basate sui dati, la disparità di trattamento

²⁰In effetti, la significatività statistica ha sempre fatto parte dei criteri EEOC insieme alla significatività sostanziale.

il toolbox è più ampio e comprende la legge sulla privacy, spiegazioni e potenzialmente la legge sulla protezione dei consumatori. Ne discutiamo a turno.

Privacy

Quando ci preoccupiamo della privacy, la preoccupazione di fondo è spesso che i dati che ci riguardano possano essere utilizzati per discriminare o comportare un trattamento sfavorevole. Ad esempio, se un intervistatore fa domande sulla religione, ciò potrebbe essere considerato una violazione della privacy. Il danno che anima questa preoccupazione è la negazione di un lavoro. Come altro esempio, i rapporti secondo cui il rivenditore Target utilizza i registri degli acquisti per identificare le clienti incinte ha suscitato indignazione.²⁷² Il rischio di danni deriva dal fatto che la gravidanza è un momento in cui gli individui sono particolarmente suscettibili alla manipolazione attraverso il marketing (che è il motivo per cui gli operatori di marketing sono interessati alla gravidanza). in primo luogo).

Tuttavia, negli Stati Uniti la legge sulla privacy dei dati e la legge antidiscriminazione sono state in gran parte separate . Tornando all'esempio precedente, non è la legge sulla privacy a vietare agli intervistatori di fare domande sulla religione. Piuttosto, poiché il diritto del lavoro vieta la discriminazione sulla base della religione, l'orientamento interpretativo dell'EEOC e, spesso, delle istituzioni stesse scoraggia tali domande

durante i colloqui.²¹ Tuttavia, dato l'allineamento normativo, è naturale chiedersi se la legge sulla privacy possa essere adattata per servire fini di antidiscriminazione. C'è molto fascino intuitivo in questa idea, soprattutto quando si tratta di apprendimento automatico. Se un sistema decisionale si basa sui dati, perché non porre restrizioni al flusso di dati per prevenire discriminazioni ingiustificate?

Ma quando esaminiamo questo argomento più nel dettaglio emergono delle difficoltà. La più ovvia è la questione delle deleghe. Come abbiamo discusso nel capitolo 3, vietare l'accesso ad attributi sensibili come la razza o il genere ha in genere un impatto trascurabile su un classificatore quando sono disponibili set di dati ricchi. Non si tratta solo del fatto che il decisore può addestrare un modello a prevedere l'attributo sensibile da attributi innocui, come nell'esempio Target sopra. Potrebbe invece utilizzare direttamente attributi innocui per prevedere l'esito dell'interesse, come la suscettibilità di una particolare persona a un particolare messaggio di marketing. Questo è infatti esattamente ciò che è stato dimostrato che accade sulle piattaforme pubblicitarie su scala Facebook.²⁷⁴

Se il problema sono i proxy, un altro approccio consiste nel vietare la raccolta di proxy. Questa è l'idea alla base delle leggi "ban the box" negli stati americani che vietano ai datori di lavoro di informarsi sui precedenti penali. Il ban-the-box ha due motivazioni.

Il primo è facilitare il reinserimento nella società degli ex detenuti attraverso il lavoro. In questa prospettiva, la stessa storia criminale può essere vista come un attributo sensibile. L'altro motivo è combattere l'impatto razziale della discriminazione contro le persone precedentemente incarcerate. Qui, la storia criminale può essere vista come un indicatore della razza. È questa visione che ci interessa.

²¹Un primo tentativo di conciliare gli obiettivi di privacy e responsabilità (ma non quelli di antidiscriminazione) è riscontrabile nei Principi sulla pratica della corretta informazione. Il FIPPS contiene le basi di leggi complete sulla protezione dei dati emanate in tutto il mondo. Negli Stati Uniti non hanno forza di legge, tranne che in alcune leggi settoriali come il Fair Credit Reporting Act. Una versione annacquata del FIPPS, incentrata su "notifica e scelta", governa il commercio online negli Stati Uniti.²⁷³

Un autorevole studio di Agan e Starr ha scoperto che i datori di lavoro aumentavano la discriminazione razziale quando erano soggetti alle leggi "ban-the-box".²⁷⁵ Cosa significa questo per la prospettiva di prevenire la discriminazione vietando i flussi di informazioni? Un punto di vista è che, alla luce di questa scoperta, le leggi ban-the-box chiaramente danneggiano più di quanto aiutano. Ma un'altra prospettiva è che la discriminazione razziale è già illegale, quindi ciò che lo studio rivela realmente è la necessità di intensificare i controlli e l'applicazione delle norme. Se ciò dovesse accadere, le leggi ban-the-box potrebbero essere in grado di raggiungere gli effetti desiderati.

Andando oltre gli attributi protetti e i loro proxy, la legge sulla privacy potrebbe rendere più difficile accumulare dossier sugli individui (ad esempio, contenenti record di acquisti o di navigazione), e potremmo sperare che ciò renda più difficile la discriminazione. Anche se una discussione approfondita di questo argomento va oltre il nostro ambito, la legge statunitense sulla privacy viene spesso criticata per non riuscire a raggiungere questo obiettivo in modo efficace, per diversi motivi. Nell'UE non esiste una legge federale generale sulla privacy analoga al GDPR. Esistono solo poche leggi settoriali sulla privacy, come l'Health Insurance Portability and Accountability Act (HIPAA). La privacy nella maggior parte delle transazioni o interazioni commerciali si riduce alla "nota e scelta", che è inefficace per molte ragioni, tra cui l'asimmetria di potere e l'asimmetria informativa tra aziende e individui. Nel contesto dell'apprendimento automatico, l'approccio di notifica e scelta alla privacy è particolarmente inefficace come barriera per le aziende che costruiscono modelli che potrebbero dedurre attributi sensibili o prendere decisioni avverse sulla base di attributi innocui. Ciò è dovuto alla "tirannia della minoranza": basta solo un piccolo numero di individui che acconsentono alla raccolta per poter scoprire i modelli statistici che rendono possibili tali inferenze.²⁷⁶ Sebbene le leggi sulla privacy non abbiano finora contribuito ad affrontare la discriminazione, la legge sulla discriminazione ha talvolta contribuito a preservare la

privacy. Il Genetic Information Nondiscrimination Act è una legge antidiscriminazione che si è trasformata in una legge sulla privacy attraverso ampie decisioni dei tribunali e l'interpretazione dell'EEOC.²⁷⁷ Le informazioni genetiche rappresentano un'eccezione all'ubiquità dei proxy, poiché non possono essere facilmente dedotte con alcun grado di completezza o accuratezza da caratteristiche osservabili.

I significati più ampi del termine privacy vanno oltre il flusso di informazioni e comprendono trasparenza, spiegazione e risarcimento. Passiamo a quelli successivi.

Spiegazione

Nel contesto del processo decisionale automatizzato, la spiegazione potrebbe avere uno di questi due obiettivi. La prima è una spiegazione del sistema complessivo. In un sistema basato su regole questo potrebbe essere l'insieme delle regole decisionali. In un sistema di apprendimento automatico è meno ovvio quale forma dovrebbe assumere questa spiegazione, ed è oggetto di ricerca attiva nel campo dell'apprendimento automatico interpretabile.

Una spiegazione del sistema complessivo promuove obiettivi di equità perché consente ai regolatori, agli utenti e agli sviluppatori di verificare se il sistema aderisce ai requisiti normativi. In molti casi, la spiegazione ci consente di individuare immediatamente potenziali ingiustizie. Ad esempio, se sappiamo che un sistema utilizzato per rilevare account falsi sui social media si basa su un nome non comune come segnale di inautenticità,

è facile capire perché potrebbe essere più probabile che vengano segnalati erroneamente gli utenti che appartengono a minoranze culturali, come abbiamo discusso nel Capitolo 1.

Il secondo obiettivo è spiegare come è stata presa una particolare decisione date le caratteristiche dell'oggetto della decisione. Questo obiettivo può anche promuovere obiettivi di equità. Soddisfa un potente bisogno innato di comprendere come vengono prese le decisioni consequenziali su di noi. Il terrore che nasce quando un sistema decisionale ci nega tale spiegazione è abbastanza viscerale da avere un nome: kafkiano. La spiegazione delle decisioni individuali serve anche a scopi più strumentali. Ci consente di contestare decisioni che potrebbero essere state prese sulla base di informazioni errate.

Anche se la decisione fosse accurata, la spiegazione consente il ricorso, cioè le azioni che i soggetti della decisione potrebbero intraprendere per modificare la decisione in futuro. Ad esempio, un richiedente di prestito a cui è stato rifiutato a causa di un punteggio di credito basso può tentare di migliorare quel punteggio.

Facendo un passo indietro, i sistemi decisionali possono essere analizzati a tre livelli. Il livello più alto è quello dei valori, degli obiettivi e dei vincoli normativi (ad esempio, massimizzare l'accuratezza predittiva garantendo al tempo stesso l'equità). Il secondo è la progettazione del sistema e le sue regole. Il terzo è il livello delle decisioni individuali. La giustificazione è necessaria a tutti e tre i livelli. Nei sistemi decisionali tradizionali, i valori e gli obiettivi ottengono legittimità attraverso la partecipazione delle parti interessate, la deliberazione e il dibattito democratico. Spesso si fonde con il passaggio successivo, la regolamentazione o la definizione delle politiche, che è il processo che porta dal primo livello al secondo, progettando un sistema decisionale basato su valori e obiettivi. Se si salta il primo passaggio, in questo processo diventano evidenti le tensioni tra valori diversi o tra obiettivi dei diversi stakeholder. Nelle burocrazie amministrative, i problemi vengono risolti attraverso un processo di partecipazione pubblica.²⁷⁸ Al contrario, il processo di aggiudicazione collega il secondo e il terzo livello.²⁷⁹ Ad esempio, negli Stati Uniti, i burocrati valutano periodicamente il valore delle case e di altri beni immobili, sulla base di una politica elaborata al fine di determinare l'importo dell'imposta sulla proprietà da imporre. Se il proprietario non è d'accordo con la valutazione, può presentare ricorso e ha diritto a un'udienza.

I sistemi automatizzati erodono le protezioni procedurali coinvolte nella regolamentazione e nell'aggiudicazione: rispettivamente partecipazione pubblica e ricorsi.²⁸⁰ Questi problemi sono esacerbati quando è coinvolto l'apprendimento automatico, a causa della sua imperscrutabilità e non intuitività.²⁸¹ I due obiettivi della spiegazione potrebbero aiutare a mitigare queste preoccupazioni: permettendoci di comprendere come il sistema complessivo e la politica si conformano ai vincoli normativi e come le decisioni individuali si conformano alla politica. Questi corrispondono grosso modo alla distinzione tra interpretabilità "globale" e "locale" nella letteratura tecnica.

I requisiti per entrambi i tipi di spiegazione possono essere visti nelle leggi esistenti. L'FCRA e l'ECOA contengono un requisito di "avviso di azioni avverse". Questo è un esempio del secondo obiettivo, poiché riguarda solo la decisione individuale e non richiede trasparenza sul modello complessivo. Al contrario, il GDPR richiede "informazioni significative sulla logica coinvolta" se un individuo è soggetto a una decisione consequenziale da parte di un sistema automatizzato. Ciò è generalmente inteso come la necessità di un certo grado di spiegazione sia del modello generale che della decisione specifica.

Tabella 6.2: Confronto tra le due tipologie di spiegazione del modello

	Spiegazione del sistema generale	Spiegazione di decisioni specifiche
Obiettivo	Giustificare la politica in base agli obiettivi alla politica dei valori	Giustificare la decisione in base
Analogico burocratico	Decisione normativa o politica	
Strumento tecnico	Interpretabilità globale	Spiegazione locale
Esempio di requisito legale	GDPR: "informazioni significative sulla logica coinvolta"	FCRA ed ECOA: negativi avviso di azione

Selbst e Barcas descrivono diverse limitazioni all'utilità delle spiegazioni.

Ne evidenziamo due principali. Il primo è la difficoltà di produrre spiegazioni che siano allo stesso tempo fedeli al modello e comprensibili a un non esperto.

Se un modello di credito combina dozzine di variabili in modi non lineari, ragioni come "durata del rapporto di lavoro" o "reddito insufficiente" potrebbero non essere sufficienti a spiegare pienamente una decisione; tuttavia questo è tutto ciò che è richiesto per gli avvisi di azioni avverse.

Al contrario, una spiegazione di una decisione che sia pienamente fedele ad un modello statistico può risultare incomprensibile alla maggior parte dei soggetti decisionali.

Esiste un'importante distinzione tra le spiegazioni fornite volontariamente e quelle richieste dalla legge a un decisore che non ha altri incentivi a fornirle . Finora, è stato difficile per le autorità di regolamentazione stabilire requisiti legali per ciò che costituisce una buona spiegazione e valutare se funzionano come previsto. L'evidenza empirica supporta la difficoltà di costringere i decisori riluttanti a fornire spiegazioni significative. Ad esempio, uno studio del 2018 ha rilevato che la domanda "Perché vedo questo?" le spiegazioni degli annunci sono vaghe, incomplete, fuorvianti e generalmente inutili.²⁸² La letteratura scientifica mostra che se Facebook volesse fornire buone spiegazioni, sarebbe possibile fare molto meglio.

Una limitazione più fondamentale descritta da Selbst e Barcas è che anche le spiegazioni fedeli e comprensibili potrebbero non consentire una valutazione normativa . Se un datore di lavoro utilizza un modello di screening che calcola un punteggio basato su alcune parole chiave (una spiegazione fedele e comprensibile), è normativamente importante sapere se tali parole chiave rappresentano competenze legate al lavoro o agiscono come proxy (ad esempio, hobby) che segnalano classe sociale o qualcos'altro. Potremmo essere in grado di fare una valutazione del genere date le parole chiave, ma non è semplice. I metodi moderni che forniscono spiegazioni basate su concetti di alto livello piuttosto che su caratteristiche di basso livello sono promettenti a questo riguardo, ma è probabile che il divario tra le spiegazioni e una piena giustificazione normativa rimanga.

A causa di queste limitazioni, si è verificato un passaggio graduale dalle spiegazioni alle valutazioni d'impatto algoritmiche (AIA). Una discussione completa delle AIA va oltre il nostro scopo, ma sottolineiamo come le AIA, almeno in una versione idealizzata, differiscono dalle spiegazioni. Innanzitutto, le AIA vanno oltre la spiegazione del modello stesso e si concentrano sul come

è stato creato, come verrà utilizzato e quali impatti potrebbe avere. In secondo luogo, i principali consumatori delle AIA non sono i soggetti decisionali ma piuttosto i regolatori e altri esperti, il che allevia in una certa misura il compromesso fedeltà-comprensibilità. In terzo luogo, le AIA devono essere eseguite prima che il modello venga implementato e devono essere aggiornate periodicamente. Alcune visioni delle AIA richiedono il coinvolgimento di soggetti esterni imparziali nella loro produzione.

Il GDPR incorpora una versione delle AIA, ovvero le valutazioni dell'impatto sulla protezione dei dati (DPIA). Le DPIA devono includere una descrizione dell'algoritmo e dello scopo del trattamento, una valutazione della necessità del trattamento in relazione allo scopo, una valutazione dei rischi per i diritti e le libertà individuali e le misure che un'azienda utilizzerà per affrontarli rischi. Richiede la consultazione "ove appropriato" con le persone colpite. Ma non è necessario che le DPIA siano rese pubbliche. È troppo presto per dire quanto saranno efficaci nella pratica; molto dipenderà dal comportamento delle autorità di regolamentazione.²⁸³,
²⁸⁴

Le valutazioni di impatto algoritmico sono strettamente correlate agli audit e i termini sono talvolta usati in modo intercambiabile. Tuttavia, ci sono diversi tipi che vale la pena distinguere. Un rapporto del 2020 li classifica in quattro categorie:²⁸⁵

- Audit sui pregiudizi condotti da ricercatori, giornalisti o organizzazioni della società civile (ispirati agli audit delle scienze sociali, come abbiamo visto nel capitolo sui test pratici sulla discriminazione).
- Audit regolamentari condotti da autorità di regolamentazione con poteri statutari per esaminare dati e sistemi interni, sul modello degli audit finanziari.
- Valutazioni algoritmiche del rischio condotte dallo sviluppatore o dal committente di uno strumento, modellate sulle valutazioni di impatto ambientale, per valutare i possibili rischi e le strategie di mitigazione prima di implementare un sistema.
- Valutazioni algoritmiche di impatto, che sono retrospettive e modellate su valutazioni politiche, condotte tipicamente da agenzie del settore pubblico rispetto ad algoritmi che implementano una politica.

Le valutazioni d'impatto e gli audit algoritmici sono un settore in piena espansione²⁸⁶ e il loro potenziale è ancora in fase di studio. Ad esempio, Ifeoma Ajunwa sostiene ambiziosamente l'introduzione nel diritto del lavoro esistente di un dovere di diligenza che obblighi i datori di lavoro a condurre controlli sui sistemi di assunzione automatizzati.²⁸⁷ Malgieri e Pasquale propongono un modello di regolamentazione ex ante in cui gli sviluppatori di sistemi di intelligenza artificiale consequenziale devono eseguire un valutazione del rischio prima dell'implementazione e, in alcuni casi, è necessario ottenerne l'approvazione da parte di un'autorità[@malgieri2022transparency]. Questi sviluppi illustrano la nostra tesi secondo cui la svolta verso l'apprendimento automatico, pur creando sfide per la legge antidiscriminatoria, crea anche opportunità . Lo strumento software e i record di dati coinvolti nei sistemi automatizzati forniscono un punto di leva per le autorità di regolamentazione.

Tutela dei consumatori

La legge sulla tutela dei consumatori ha radici completamente diverse sia dalla legge antidiscriminatoria che dalla legge sulla privacy. I movimenti dei consumatori hanno guadagnato terreno per la prima volta nel

Stati Uniti all'inizio del XX secolo, inizialmente a causa di problemi di sicurezza alimentare.²⁸⁸ La Federal Trade Commission è stata istituita nel 1914. Sebbene inizialmente si concentrasse sull'antitrust, la protezione dei consumatori divenne gradualmente un polo altrettanto importante delle sue attività. È stata la principale agenzia responsabile della protezione dei consumatori e ha l'autorità statutaria per contrastare le pratiche "sleali o ingannevoli" nel commercio. È questa autorità che l'agenzia utilizza per svolgere le attività per le quali è ben nota, come il controllo della pubblicità ingannevole e delle frodi, in particolare il furto d'identità.²⁸⁹ Molti stati hanno leggi sulla protezione dei consumatori di portata simile, applicate dai procuratori generali.

La regolamentazione del credito è un ambito del diritto a tutela dei consumatori che serve anche a fini di equità, intesa in senso lato. Il Fair Credit Reporting Act del 1974 restringe gli usi consentiti dei rapporti di credito in modo che non vengano utilizzati per scopi arbitrari. Offre ai consumatori la possibilità di contestare le inesattezze nei dati, considerando che vengono utilizzati per prendere decisioni consequenziali. E richiede di informare il consumatore quando vengono intraprese azioni negative nei suoi confronti. L'FCRA non affronta la discriminazione nel senso di trattamento disparato o impatto disparato; ciò sarebbe avvenuto più tardi, con l'Equal Credit Opportunity Act del 1976. In altri settori, come il diritto del lavoro, la protezione dei consumatori attualmente non svolge alcun ruolo, sebbene gli studiosi abbiano speculativamente sostenuto di trattare i candidati al lavoro come consumatori.²⁹⁰ Al di fuori

dei settori tradizionali di antidiscriminazione legge, esiste una vasta gamma di prodotti digitali di uso quotidiano in cui si manifestano pregiudizi legati all'apprendimento automatico, ed è qui che la legge sulla protezione dei consumatori è potenzialmente molto rilevante. Ad esempio, se una funzione di sblocco facciale su uno smartphone è sostanzialmente meno precisa per alcuni gruppi di utenti, ciò non costituisce una violazione di nessuno degli statuti specifici del settore di cui abbiamo discusso finora, ma potrebbe rientrare nell'autorità della FTC. Anche in settori come la discriminazione sul lavoro, esistono lacune peculiari, come il fatto che i fornitori di strumenti di screening algoritmico non sono entità coperte, e la legislazione sulla tutela dei consumatori può potenzialmente aiutare a colmare questa lacuna.

Al momento della stesura di questo articolo, tutto questo è speculativo. Finora, la FTC non ha perseguito pratiche discriminatorie, tranne quando l'azienda viola anche uno statuto antidiscriminatorio come l'ECOA, che la FTC ha l'autorità di far rispettare.²⁹⁰ Il termine "ingiustizia" nella legge della FTC è stato utilizzato tradizionalmente significa qualcosa di completamente diverso: trarre vantaggio ingiustificato dai consumatori che non possono evitare. L'esempio tipico di una pratica commerciale sleale sarebbe la vendita di olio di serpente, mentre un esempio più moderno sarebbe una scarsa sicurezza dei dati che porta a una violazione dei dati. Ma si noti che, a differenza degli statuti antidiscriminazione, la FTC ha sostanzialmente più potere nel determinare cosa costituisce ingannevole e ingiusto. È del tutto possibile che l'agenzia adotti una visione ampia dell'ingiustizia e che i tribunali lo consentano. Lo statuto consente alla FTC di guardare alle "politiche pubbliche consolidate" per determinare cosa è ingiusto. È stato suggerito che la FTC possa quindi considerare gli statuti e le regole antidiscriminazione come un'impalcatura per costruire un quadro per prendere decisioni sulla discriminazione algoritmica.²⁹¹

L'autorità ingannevole della FTC offre un'opzione più chiara ma più circoscritta. Le aziende spesso affermano affermativamente che i loro prodotti sono imparziali. Se tali affermazioni si rivelano false, è un inganno. Lo stesso vale per il falso

affermazioni sull'efficacia dei prodotti. Ciò è rilevante poiché molti strumenti decisionali predittivi presenti sul mercato non dispongono di prove di validità predittiva, il che significa che potrebbero sottoporre le persone a decisioni arbitrarie. Tuttavia, a meno che tali decisioni arbitrarie non siano anche sistematicamente distorte, sono difficili da contestare ai sensi della legge antidiscriminatoria. Infine, anche la mancanza di trasparenza può costituire una pratica ingannevole. In effetti, la FTC è intervenuta contro un'azienda che ha addestrato un modello di riconoscimento facciale sulle foto dei suoi utenti, dicendo loro falsamente che la funzione era attivabile.²⁹² Storicamente, la FTC ha avuto un percorso sulle montagne russe in termini di ampiezza tratta la sua

autorità e quanto flette i muscoli. Dopo essere stato inefficace negli anni '60 e rinvigorito negli anni '70,²⁹³ il Congresso lo ha rimproverato all'inizio degli anni '80 e ha limitato la sua autorità a causa delle pressioni esercitate da potenti interessi economici.²⁹⁴ Da allora è rimasto cauto, ed è stato ulteriormente colto di sorpresa nel era tecnologica a causa delle limitazioni delle competenze tecniche interne. Ciò ha portato a critiche accanite per fallimenti come quello di aver consentito l'esfiltrazione dei dati degli utenti di Facebook da parte di Cambridge Analytica , nonostante la FTC fosse a conoscenza da tempo di eventi precedenti simili e presumibilmente monitorasse da vicino Facebook nell'ambito di un "decreto di consenso". Negli anni '20, l'agenzia ha mostrato alcuni segni di rinvigorimento. Nello specifico sulla discriminazione algoritmica, ha pubblicato un post sul blog contenente un linguaggio sorprendentemente forte.²⁹⁵ Anche un libro bianco , scritto in collaborazione con un commissario in carica, delinea un programma ambizioso.²⁹⁶ Tutto questo per dire: la rilevanza della legge sulla tutela dei consumatori per la discriminazione algoritmica rimane una questione jolly.

Al di là della concezione tradizionale di protezione dei consumatori, stanno emergendo idee come un dovere di lealtà per le aziende a cui vengono affidati i dati dei clienti.²⁹⁶ Tali aziende sarebbero obbligate ad agire nel migliore interesse delle persone che espongono i propri dati e le proprie esperienze online. Il dovere di lealtà è un obbligo comune nei rapporti fiduciari (ad esempio, un avvocato ha tale dovere nei confronti del suo cliente). Ma la sua applicazione ai titolari di dati personali è un'idea relativamente nuova. Sebbene sia stato proposto principalmente con l'obiettivo di migliorare la privacy e ridurre al minimo le pratiche manipolative come i "dark pattern", avrebbe alcune implicazioni anche per la non discriminazione.

Considerazioni conclusive

Abbiamo trattato molti argomenti in questo capitolo. Abbiamo esaminato come i vari movimenti per i diritti civili abbiano dato origine, insieme, a un corpus di leggi antidiscriminazione relativamente robusto negli Stati Uniti. In generale, questa legge mira a trovare un equilibrio tra la prevenzione (e il rimedio) della discriminazione, dall'altro, e l'evitamento di un'eccessiva discriminazione . oneri per i decisor, dall'altro. È stato perfezionato, contestato e implementato nel corso di decenni grazie al push e al pull delle decisioni dei tribunali, delle agenzie di regolamentazione, dei burocrati istituzionali, del continuo attivismo per i diritti civili e dei cambiamenti nell'opinione pubblica. Presenta importanti limiti: in pratica, i ricorrenti privati hanno difficoltà a trovare ricorso legale; più fondamentalmente, la legge stessa è lungi dall'essere un percorso ideale per apportare cambiamenti strutturali.

Passando alle nuove sfide sollevate dal processo decisionale automatizzato, esiste il rischio che l'apprendimento automatico discriminatorio possa colmare le lacune nel modo in cui la legge concepisce la discriminazione. A nostro avviso, questo rischio è controbilanciato dall'ampio pacchetto di strumenti legali disponibili: legge sulla privacy, requisiti relativi alla spiegazione e alla valutazione dell'impatto e legge sulla protezione dei consumatori. Finora, questo potenziale è rimasto per lo più dormiente per vari motivi: una concezione ristretta della privacy, la mancanza di una legislazione ampia negli Stati Uniti che richieda una spiegazione delle decisioni consequenziali e la timidezza delle agenzie di tutela dei consumatori. Ciò potrebbe ancora cambiare; è possibile che la legge e le forze dell'ordine possano essere riformate per affrontare efficacemente i nuovi problemi. Come minimo, anche se non sanciti dalla legge, gli strumenti di intervento di cui abbiamo discusso offrono un modello per i difensori dell'interesse pubblico che cercano di ritenere le aziende responsabili.

7

Testare la discriminazione nella pratica

Nei capitoli precedenti abbiamo visto i criteri di equità statistica, causale e normativa. Questo capitolo tratta delle complessità che sorgono quando vogliamo applicarli nella pratica.

Un tema ricorrente di questo libro è che non esiste un unico test per l'equità, cioè non esiste un unico criterio che sia allo stesso tempo necessario e sufficiente per l'equità.

Piuttosto, esistono molti criteri che possono essere utilizzati per diagnosticare potenziali ingiustizie o discriminazioni.

C'è spesso un divario tra le nozioni morali di equità e ciò che è misurabile con i metodi sperimentali o osservativi disponibili. Ciò non significa che possiamo selezionare e applicare un test di equità basato sulla convenienza. Lungi da ciò: abbiamo bisogno di un ragionamento morale e di considerazioni specifiche per il dominio per determinare quali test sono appropriati, come applicarli, determinare se i risultati indicano una discriminazione ingiusta e se è necessario un intervento. Vedremo esempi di tale ragionamento in questo capitolo. Al contrario, se un sistema supera un test di equità, non dovremmo interpretarlo come un certificato che il sistema è giusto.

In questo capitolo, i nostri principali oggetti di studio saranno i sistemi reali piuttosto che i modelli di sistemi. Dobbiamo tenere presente che ci sono molti presupposti necessari nella creazione di un modello che potrebbero non essere validi nella pratica. Ad esempio, i cosiddetti sistemi decisionali automatizzati raramente operano senza alcun giudizio umano.

Oppure possiamo supporre che un sistema di apprendimento automatico venga addestrato su un campione tratto dalla stessa popolazione su cui prende decisioni, il che non è quasi mai vero nella pratica. Inoltre, il processo decisionale nella vita reale raramente è costituito da un singolo punto decisionale, ma piuttosto da una serie cumulativa di piccole decisioni. Ad esempio, l'assunzione comprende l'approvvigionamento, lo screening, i colloqui, la selezione e la valutazione, e queste stesse fasi includono molte componenti.²⁹⁷ Un'importante

fondamentale fonte di difficoltà per testare la discriminazione nella pratica è che i ricercatori hanno una capacità limitata di osservare – e ancor meno di manipolare. - molti dei passaggi in un sistema del mondo reale. Vedremo infatti che anche il decisore si trova ad affrontare limiti nella sua capacità di studiare il sistema.

Nonostante queste limitazioni e difficoltà, testare empiricamente l'equità è vitale. Gli studi di cui parleremo servono come prova dell'esistenza della discriminazione e forniscono un limite inferiore della sua prevalenza. Consentono di monitorare le tendenze della discriminazione nel tempo. Quando i risultati sono sufficientemente evidenti, ne giustificano la necessità

intervento indipendentemente da eventuali differenze interpretative. E quando applichiamo un intervento di equità, ci aiutano a misurarne l'efficacia. Infine, la ricerca empirica può anche aiutare a scoprire i meccanismi attraverso i quali avviene la discriminazione, consentendo interventi più mirati ed efficaci. Ciò richiede un'attenta formulazione e verifica delle ipotesi utilizzando la conoscenza del dominio.

La prima metà di questo capitolo esamina i classici test di discriminazione che sono stati sviluppati nel contesto dei sistemi decisionali umani. I concetti sottostanti sono altrettanto applicabili allo studio dell'equità nei sistemi automatizzati. Gran parte della prima metà si baserà sul capitolo sulla causalità e spiegherà tecniche concrete tra cui esperimenti, differenza nelle differenze e discontinuità di regressione. Sebbene questi siano strumenti standard nel toolkit dell'inferenza causale, impareremo i modi specifici in cui possono essere applicati alle questioni di equità. Si passerà poi all'applicazione dei criteri osservativi del capitolo 3. La tabella riepilogativa alla fine del primo semestre elenca, per ciascun test, il criterio di equità che sonda, il tipo di accesso al sistema richiesto e altre sfumature e limitazioni. La seconda metà del capitolo riguarda la verifica dell'equità nel processo decisionale algoritmico, concentrandosi su questioni specifiche dei sistemi algoritmici.

Due rapidi punti sulla terminologia: useremo i termini ingiustizia e discriminazione più o meno come sinonimi. Non esiste una definizione generale di nessuno dei due termini, ma renderemo la nostra discussione precisa facendo riferimento a un criterio specifico quando possibile. Utilizzeremo "sistema" come abbreviazione per un sistema decisionale, come l'assunzione in un'azienda. Può comportare o meno l'automazione o l'apprendimento automatico.

Parte 1: Test tradizionali di discriminazione

Studi di audit

Lo studio di audit è una tecnica popolare per diagnosticare la discriminazione. Si tratta di un disegno di studio chiamato esperimento sul campo. "Campo" si riferisce al fatto che si tratta di un esperimento sull'effettivo sistema decisionale di interesse (sul "campo", in contrapposizione a una simulazione di laboratorio del processo decisionale). Gli esperimenti su sistemi reali sono difficili da realizzare. Ad esempio, di solito dobbiamo tenere i partecipanti all'oscuro del fatto che stanno partecipando a un esperimento. Ma gli esperimenti sul campo ci permettono di studiare il processo decisionale così come avviene realmente, invece di preoccuparci che ciò che stiamo scoprendo sia un artefatto di un ambiente di laboratorio. Allo stesso tempo, l'esperimento, manipolando e controllando attentamente le variabili, ci consente di osservare un effetto del trattamento, piuttosto che limitarsi ad osservare una correlazione.

Come interpretare un simile effetto terapeutico è una questione più complicata. A nostro avviso, la maggior parte degli studi di audit, compresi quelli che descriveremo, sono meglio visti come tentativi di testare la cecità: se un decisore utilizza direttamente un attributo sensibile. Ricordiamo che questa nozione di discriminazione non è necessariamente controfattuale in un modello causale valido (Capitolo 5). Anche nel caso dei test di cecità, si discute esattamente su cosa misurano, dal momento che il ricercatore può, nella migliore delle ipotesi, segnalare razza, genere,

o un altro attributo sensibile. Ciò diventerà chiaro quando discuteremo di studi specifici.

Gli studi di audit sono stati avviati dal Dipartimento statunitense per l'edilizia abitativa e lo sviluppo urbano negli anni '70 con lo scopo di studiare il trattamento sfavorevole subito dagli acquirenti e dagli affittuari di case appartenenti a minoranze.²⁹⁸ Da allora sono stati applicati con successo a molti altri ambiti.

In uno studio fondamentale di Ayres & Siegelman, i ricercatori hanno reclutato 38 tester per visitare circa 150 concessionarie di automobili per contrattare le auto e registrare il prezzo offerto loro alla fine della contrattazione.²⁹⁹ I tester hanno visitato le concessionarie in coppia; i tester in una coppia differivano in termini di razza o sesso. Entrambi i tester in coppia contrattavano per lo stesso modello di auto, presso la stessa concessionaria, di solito a pochi giorni di distanza l'uno dall'altro.

Portare a termine un esperimento come questo in modo convincente richiede un'attenta attenzione ai dettagli; qui descriviamo solo alcuni dei tanti dettagli presenti nel documento. Ancora più significativo, i ricercatori hanno fatto di tutto per ridurre al minimo eventuali differenze tra i tester che potrebbero essere correlate alla razza o al genere. In particolare, tutti i partecipanti al test avevano un'età compresa tra 28 e 32 anni, avevano 3-4 anni di istruzione post-secondaria e "sono stati scelti soggettivamente per avere un'attrattiva media". Inoltre, per ridurre al minimo il rischio che l'interazione dei tester con i rivenditori fosse correlata alla razza o al genere, ogni aspetto del loro comportamento verbale o non verbale era governato da un copione. Ad esempio, tutti i tester "indossavano abiti sportivi simili 'yuppie' e si recavano al concessionario con auto a noleggio simili". Hanno anche dovuto memorizzare le risposte a un lungo elenco di domande che probabilmente avrebbero incontrato. Tutto ciò ha richiesto una formazione approfondita e debrief regolari.

La scoperta principale del documento è stata una penalità di prezzo ampia e statisticamente significativa nelle offerte ricevute dai tester neri. Ad esempio, i maschi neri hanno ricevuto offerte finali che erano circa 1.100 dollari in più rispetto ai maschi bianchi, il che rappresenta una differenza tripla nei profitti del rivenditore sulla base dei dati sui costi del rivenditore. L'analisi nel documento prevede variabili target alternative (offerte iniziali invece di offerte finali; margine percentuale invece di offerte in dollari), specifiche di modelli alternativi (ad esempio per tenere conto dei due audit in ciascuna coppia con rumore correlato) e controlli aggiuntivi (ad esempio strategia di contrattazione). Pertanto, ci sono una serie di stime diverse, ma i risultati principali rimangono solidi.¹ Un'interpretazione allentante di questo studio è che se due persone fossero identiche

tranne che per la razza, con una bianca e l'altra nera, allora le offerte che dovrebbero aspettarsi di ricevere differirebbe di circa \$ 1.100. Ma cosa significa che due persone sono identiche tranne che per la razza? Quali attributi su di loro sarebbero gli stessi e quali sarebbero diversi?

Grazie alla discussione sull'instabilità ontologica nel capitolo 5, possiamo comprendere la struttura implicita degli autori per prendere queste decisioni. A nostro avviso, trattano la razza come un nodo sorgente stabile in un grafico causale, che tentano di mantenere

¹In un esperimento come questo in cui il trattamento è randomizzato, l'aggiunta o l'omissione di variabili di controllo in una stima di regressione dell'effetto del trattamento non risulta in una stima errata, ma le variabili di controllo possono spiegare parte del rumore nelle osservazioni e quindi aumentare la precisione della stima dell'effetto del trattamento, ovvero diminuire l'errore standard del coefficiente.

costante tutti i suoi discendenti, come abbigliamento e comportamento, al fine di stimare l' effetto diretto della razza sul risultato. Ma cosa succederebbe se uno dei meccanismi di ciò che intendiamo come "discriminazione razziale" si basasse su differenze di abbigliamento e comportamento?

La costruzione sociale della razza suggerisce che ciò sia plausibile.³⁰⁰

Si noti che gli autori non hanno tentato di eliminare le differenze di accento tra i tester. Perché no? Da un punto di vista pratico, l'accento è difficile da manipolare.

Ma una difesa più fondata sulla scelta degli autori è che l'accento fa parte del modo in cui comprendiamo la razza; una parte di ciò che significa essere nero, bianco, ecc., in modo che anche se i tester potessero manipolare i loro accenti, non dovrebbero. L'accento è incluso nel nodo "razza" nel grafico causale.

Per assumere una posizione informata su questioni come questa, abbiamo bisogno di una profonda comprensione del contesto culturale e della storia. Sono oggetto di un acceso dibattito in sociologia e nella teoria critica della razza. Il nostro punto è questo: la progettazione e l'interpretazione degli studi di audit richiedono l'assunzione di posizioni su questioni sociali controverse. Potrebbe essere inutile cercare un unico modo "corretto" per testare anche il concetto apparentemente semplice di equità secondo cui il decisore tratta individui simili allo stesso modo, indipendentemente dalla razza. Controllare una plethora di attributi è un approccio. Probabilmente, fornisce limiti più bassi alla quantità di discriminazione poiché incorpora una concezione ristretta di razza. Un altro è semplicemente reclutare tester neri e tester bianchi, farli comportare e contrattare come sarebbe la loro naturale inclinazione e misurare la disparità demografica. Ciascun approccio ci dice qualcosa di prezioso e nessuno dei due è "migliore".²

Un altro famoso studio di Bertrand & Mullainathan ha testato la discriminazione nel mercato del lavoro.³⁰¹ Invece di inviare i tester di persona, i ricercatori hanno inviato curriculum finti in risposta agli annunci di lavoro. Il loro obiettivo era verificare se la razza del candidato avesse un impatto sulla probabilità che un datore di lavoro lo invitasse per un colloquio. Hanno segnalato la razza nei curriculum utilizzando nomi dal suono bianco (Emily, Greg) o nomi dal suono nero (Lakisha, Jamal). Creando coppie di curriculum identici tranne che per il nome, hanno scoperto che i nomi bianchi avevano il 50% in più di probabilità di risultare in una richiamata rispetto ai nomi neri. L'entità dell'effetto equivaleva ad ulteriori otto anni di esperienza su un curriculum.

Nonostante l'attenta progettazione dello studio, sono inevitabilmente sorti dibattiti sull'interpretazione , principalmente a causa dell'uso dei nomi dei candidati come un modo per segnalare la razza ai datori di lavoro. I datori di lavoro hanno notato i nomi in tutti i casi e l'effetto sarebbe stato ancora più forte se l'avessero fatto? Oppure, le disparità osservate possono essere meglio spiegate sulla base di fattori correlati alla razza, come la preferenza per nomi più comuni e familiari, o una deduzione di uno status socioeconomico più elevato per i candidati con nomi dal suono bianco? (Naturalmente, le spiegazioni alternative non rendono il comportamento osservato moralmente accettabile, ma sono importanti da considerare.) Sebbene gli autori forniscano prove contro queste interpretazioni, il dibattito è continuato. Per una discussione delle critiche alla validità degli studi di audit, vedere l'indagine di Devah Pager.³⁰²

² Nella maggior parte degli altri ambiti, ad esempio l'occupazione, testare la disparità demografica avrebbe meno valore, perché esistono differenze rilevanti tra i candidati. La discriminazione di prezzo è insolita in quanto non esistono qualità moralmente salienti degli acquirenti che possano giustificarla.

In ogni caso, come altri studi di audit, questo esperimento mette alla prova l'equità come cecità. Anche i semplici proxy di razza, come il quartiere residenziale, sono stati mantenuti costanti tra coppie di curriculum abbinate. Pertanto, il progetto probabilmente sottostima la misura in cui le caratteristiche moralmente irrilevanti influenzano nella pratica i tassi di richiamata. Questo è solo un altro modo per dire che l'inversione degli attributi generalmente non produce controllati che ci interessano, e non è chiaro se le dimensioni degli effetti misurati abbiano un'interpretazione significativa che si generalizzi oltre il contesto dell'esperimento.

Piuttosto, come sostiene Issa Kohler-Hausmann, gli studi di audit sono preziosi perché innescano un'intuizione morale forte e valida.³⁰³ Hanno anche uno scopo pratico: se ben progettati, mettono in luce i meccanismi che producono disparità e aiutano a guidare gli interventi. Ad esempio, lo studio sulla contrattazione automobilistica ha concluso che le preferenze dei proprietari di concessionari non spiegano la discriminazione osservata, che le preferenze di altri clienti possono spiegarne in parte, ed è evidente che i concessionari stessi (piuttosto che proprietari o clienti) sono la fonte primaria della discriminazione osservata.

Gli studi di audit basati sui curriculum, noti anche come studi per corrispondenza, sono stati ampiamente replicati. Presentiamo brevemente alcuni risultati importanti, con l'avvertenza che potrebbero esserci errori di pubblicazione. Ad esempio, gli studi che non trovano prove di un effetto hanno in generale meno probabilità di essere pubblicati. In alternativa, i risultati nulli pubblicati potrebbero riflettere una progettazione inadeguata dell'esperimento o potrebbero semplicemente indicare che la discriminazione è espressa solo in determinati contesti.

Un'indagine del 2016 condotta da Bertrand e Duflo elenca 30 studi provenienti da 15 paesi che coprono quasi tutti i continenti, rivelando una discriminazione pervasiva contro le minoranze razziali ed etniche.³⁰⁴ Il metodo è stato utilizzato anche per studiare la discriminazione basata sul genere, sull'orientamento sessuale e sull'aspetto fisico.³⁰⁴ È stato mostrato che sono stati utilizzati anche al di fuori del mercato del lavoro, nel commercio al dettaglio e nel mondo accademico.³⁰⁴ Infine, sono state studiate le tendenze nel tempo: una meta-analisi non ha rilevato alcun cambiamento nella discriminazione razziale nelle assunzioni contro gli afroamericani dal 1989 al 2015. Sono stati rilevati alcuni segnali di diminuzione della discriminazione contro i latino-americani, anche se i dati su questa questione

erano scarsi.³⁰⁵ Nel complesso, gli studi di audit hanno contribuito a spostare il dibattito accademico e politico lontano dall'idea ingenua secondo cui la discriminazione è una preoccupazione di un'epoca passata. Da un punto di vista metodologico, il principale risultato che otteniamo dalla discussione sugli studi di audit è la complessità di definire e testare la cecità.

Testare l'impatto dell'accecamento

In alcune situazioni, non è possibile testare la cecità randomizzando la percezione di razza, genere o altri attributi sensibili da parte del decisore. Ad esempio, supponiamo di voler verificare se esiste una discriminazione di genere nella revisione tra pari in un particolare campo di ricerca. L'invio di documenti reali con identità di autore fittizio può far sì che il revisore tenti di cercare l'autore e si renda conto dell'inganno. Un disegno in cui il ricercatore cambia i nomi degli autori con quelli reali

persone è ancora più problematico.

Esiste una strategia leggermente diversa che è più praticabile: un editore di una rivista accademica nel campo della ricerca potrebbe condurre un esperimento in cui ogni articolo ricevuto viene assegnato in modo casuale per essere rivisto in cieco (in cui le identità dell'autore sono note ai revisori) o in doppio cieco (in cui le identità degli autori vengono nascoste ai revisori). In effetti, tali esperimenti sono stati condotti,³⁰⁶ ma in generale anche questa strategia può rivelarsi impraticabile.

In ogni caso, supponiamo che un ricercatore abbia accesso solo ai dati osservativi sulle politiche di revisione delle riviste e alle statistiche sugli articoli pubblicati. Tra dieci riviste nel campo della ricerca, alcune hanno introdotto la revisione in doppio cieco, e lo hanno fatto in anni diversi. Il ricercatore osserva che in ciascun caso, subito dopo il passaggio, la percentuale di articoli scritti da donne è aumentata, mentre non vi è stato alcun cambiamento per le riviste che hanno mantenuto la revisione in cieco. Sotto determinati presupposti, ciò consente di stimare l'impatto della revisione in doppio cieco sulla frazione di articoli accettati di cui sono autori donne. Questo esempio ipotetico illustra l'idea di un "esperimento naturale", così chiamato perché le condizioni simili a un esperimento sorgono a causa della variazione naturale. Nello specifico, il disegno dello studio in questo caso è chiamato differenze nelle differenze. La prima "differenza" è tra la revisione in singolo cieco e quella in doppio cieco, e la seconda "differenza" è tra le riviste (riga 2 nella tabella riepilogativa).

Le differenze nelle differenze sono metodologicamente sfumate e una trattazione completa esula dai nostri scopi.³⁰⁷ Notiamo brevemente alcune insidie. Potrebbero esserci fattori confondenti non osservati: forse il passaggio alla revisione in doppio cieco in ciascuna rivista è avvenuto come risultato di un cambiamento nella direzione editoriale, e i nuovi redattori hanno anche istituito politiche che incoraggiavano le autrici a presentare articoli forti. Potrebbero anche esserci effetti di ricaduta (che violano l'ipotesi di valore del trattamento dell'unità stabile): un cambiamento nella politica di una rivista può causare un cambiamento nell'insieme di articoli presentati ad altre riviste. I risultati sono serialmente correlati (se c'è una fluttuazione casuale nella composizione di genere del campo di ricerca a causa dell'ingresso o dell'esodo di alcuni ricercatori, l'effetto durerà molti anni). Ciò complica il calcolo dell'errore standard della stima.³⁰⁸ Infine, l'effetto del doppio cieco sulla probabilità di accettazione di articoli scritti da donne (piuttosto che sulla frazione di articoli accettati di autori donne) non è identificabile utilizzando questa tecnica senza ulteriori presupposti o controlli.

Anche se testare l'impatto dell'accecamento sembra simile al testare la cecità, c'è una differenza concettuale e pratica cruciale. Poiché non ci poniamo domande sull'impatto della razza, del genere o di un altro attributo sensibile, evitiamo di incorrere in un'instabilità ontologica. Il ricercatore non ha bisogno di intervenire sulle caratteristiche osservabili costruendo curriculum finti o addestrando i tester a utilizzare uno script di contrattazione. Invece, la variazione naturale delle caratteristiche rimane invariata; lo studio coinvolge soggetti decisionali reali. Il ricercatore interviene solo sul processo decisionale (o sfrutta la variazione naturale) e valuta l'impatto di tale intervento su gruppi di candidati definiti dall'attributo sensibile A.

Pertanto, A non è un nodo in un grafico causale, ma semplicemente un modo per dividere le unità in gruppi per l'analisi. Domande sul fatto che il decisore abbia effettivamente dedotto

l'attributo sensibile o semplicemente una caratteristica ad esso correlata sono irrilevanti per l'interpretazione dello studio. Inoltre, le dimensioni dell'effetto misurate hanno un significato che si generalizza a scenari al di là dell'esperimento. Ad esempio, uno studio ha testato l'effetto del "resume whitening", in cui i candidati appartenenti a minoranze nascondono deliberatamente elementi della loro identità razziale o etnica nei materiali delle domande di lavoro per aumentare le loro possibilità di essere richiamati.³⁰⁹ Gli effetti riportati nello studio sono significativi per persone in cerca di lavoro che si impegnano in questa pratica.

Rivelare fattori estranei alle decisioni

A volte gli esperimenti naturali possono essere utilizzati per mostrare l'arbitrarietà del processo decisionale piuttosto che l'ingiustizia nel senso di non cecità (riga 3 nella tabella riassuntiva). Ricordiamo che l'arbitrarietà è un tipo di ingiustizia di cui ci occupiamo in questo libro (Capitolo 2). L'arbitrarietà può riferirsi alla mancanza di una procedura decisionale uniforme o all'intrusione di fattori irrilevanti nella procedura.

Ad esempio, uno studio ha esaminato le decisioni prese dai giudici dei tribunali minorili della Louisiana, inclusa la durata delle sentenze.³¹⁰ È emerso che nella settimana successiva alla sconvolgente sconfitta subita dalla squadra di football della Louisiana State University (LSU), i giudici hanno imposto condanne pari al 7% in media più a lungo. L'impatto è stato maggiore per gli imputati neri. L'effetto è stato determinato interamente da giudici che hanno conseguito la laurea presso la LSU, suggerendo che l'effetto è dovuto all'impatto emotivo della perdita. Per i lettori che non hanno familiarità con la cultura del football universitario negli Stati Uniti, il giornale rileva utilmente che "Descrivere il football della LSU solo come un evento sarebbe un enorme eufemismo per i residenti dello stato della Louisiana".

Un altro noto studio di Danziger et al. sulla presunta inaffidabilità delle decisioni giudiziarie è infatti un manifesto del pericolo di confondere le variabili negli esperimenti naturali. Lo studio ha testato la relazione tra l'ordine in cui i giudici esaminano i casi di libertà condizionale e gli esiti di tali casi.³¹¹ Ha scoperto che la percentuale di sentenze favorevoli iniziava a circa il 65% all'inizio della giornata prima di scendere gradualmente fino a quasi zero subito prima della pausa pranzo dei giudici, è tornata al ~65% dopo la pausa, con lo stesso schema ripetuto per la pausa pranzo successiva!

Gli autori hanno suggerito che le risorse mentali dei giudici si esauriscono nel corso di una sessione, portando a decisioni più inadeguate. Divenne rapidamente noto come lo studio dei "giudici affamati" ed è stato ampiamente citato come un esempio della fallibilità dei decisori umani.

La scoperta sarebbe straordinaria se l'ordine dei casi fosse davvero casuale. In effetti, sarebbe così straordinario che si sostenesse che lo studio dovesse essere respinto semplicemente sulla base del fatto che la dimensione dell'effetto osservata è troppo grande per essere causata da fenomeni psicologici come l'attenzione dei giudici.³¹²

Gli autori erano ben consapevoli che l'ordine non era casuale e hanno eseguito alcuni test per vedere se era associato a fattori pertinenti al caso (poiché tali fattori potrebbero anche influenzare in modo legittimo la probabilità di un esito favorevole). Non hanno trovato tali fattori. Ma si è scoperto che non sembravano abbastanza attenti. Un'indagine di follow-up ha rivelato molteplici fattori confondenti e potenziali confondenti,

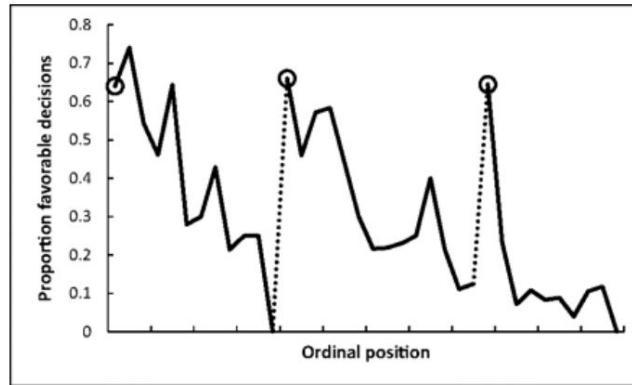


Figura 7.1: Frazione di sentenze favorevoli nel corso di una giornata. Le linee tratteggiate indicano le pause pranzo. Da Danziger et al.

compreso il fatto che i prigionieri senza avvocato vengono presentati per ultimi in ciascuna sessione e tendono a prevalere a un tasso molto più basso.³¹³ Ciò invalida la conclusione dello studio originale.

Testare l'impatto delle decisioni e degli interventi

Un aspetto sottovalutato dell'equità nel processo decisionale è l'impatto della decisione sull'oggetto della decisione. Nel nostro quadro di previsione, la variabile target (Y) non è influenzata dal punteggio o dalla previsione (R). Ma questo non è vero nella pratica.

Le banche fissano i tassi di interesse per i prestiti in base al rischio di default previsto, ma fissare un tasso di interesse più elevato aumenta le probabilità di insolvenza del mutuatario.

L'impatto della decisione sul risultato è una questione di inferenza causale.

Ci sono altre domande importanti che possiamo porci riguardo all'impatto delle decisioni. Qual è l'utilità o il costo di una decisione positiva o negativa per i diversi soggetti (e gruppi) decisionali? Ad esempio, l'ammissione a un college può avere un'utilità diversa per i diversi candidati in base agli altri college in cui sono stati o non sono stati ammessi. Le decisioni possono avere effetti anche su persone che non sono soggetti decisionali. Ad esempio, la carcerazione ha un impatto non solo sugli individui ma sulle comunità.¹⁶⁹ Misurare questi costi ci consente di essere più scientifici nel fissare soglie decisionali e nel regolare il compromesso tra falsi positivi e negativi nei sistemi decisionali.

Un modo per misurare l'impatto delle decisioni è tramite esperimenti, ma, ancora una volta, possono essere irrealizzabili per ragioni legali, etiche e tecniche. Invece, evidenziamo un disegno sperimentale naturale per testare l'impatto di una decisione – o di un intervento di equità – sui candidati, chiamato discontinuità di regressione (riga 4 nella tabella riepilogativa).

Supponiamo di voler verificare se un programma di borse di studio basato sul merito per gli studenti universitari di prima generazione ha effetti benefici duraturi, ad esempio su quanto guadagnano dopo il college. Non possiamo semplicemente confrontare lo stipendio medio degli studenti

ammaccature che hanno vinto e non hanno vinto la borsa di studio, poiché queste due variabili possono essere confuse da abilità intrinseche o altri fattori. Ma supponiamo che le borse di studio siano state assegnate in base ai punteggi dei test, con un limite dell'85%. Possiamo quindi confrontare lo stipendio degli studenti con punteggi compresi tra 85% e 86% (e quindi hanno ottenuto la borsa di studio) con quelli degli studenti con punteggi compresi tra 84% e 85% (e quindi non hanno ottenuto la borsa di studio). Possiamo supporre che all'interno di questo ristretto intervallo di punteggi dei test, le borse di studio vengano assegnate essenzialmente in modo casuale. Ad esempio, se la variazione (errore standard) nei punteggi dei test per studenti con abilità identiche è di 5 punti percentuali, la differenza tra 84% e 86% ha una significatività minima. Pertanto possiamo stimare l'impatto della borsa di studio come se avessimo effettuato uno studio randomizzato e controllato.

Dobbiamo stare attenti, però. Se consideriamo una fascia troppo ristretta di punteggi dei test attorno alla soglia, potremmo ritrovarci con dati insufficienti per l'inferenza. Se consideriamo una fascia più ampia di punteggi dei test, gli studenti di questa fascia potrebbero non essere più unità scambiabili per l'analisi.

Un'altra trappola si presenta perché si presuppone che l'insieme degli studenti che ricevono la borsa di studio siano proprio quelli sopra la soglia. Se questa ipotesi fallisce, introduce immediatamente la possibilità di fattori confondenti. Forse il punteggio del test non è l'unico criterio per la borsa di studio e il reddito viene utilizzato come criterio secondario. Oppure, alcuni studenti a cui è stata offerta la borsa di studio potrebbero rifiutarla perché hanno già ricevuto un'altra borsa di studio. Altri studenti potrebbero non avvalersi dell'offerta perché la documentazione necessaria per richiederla è ingombrante. Se è possibile sostenere il test più volte, è più probabile che gli studenti più ricchi lo facciano finché non raggiungono la soglia di ammissibilità.

Test puramente osservativi

L'ultima categoria di test quantitativi per la discriminazione è puramente osservativa. Quando non siamo in grado di fare esperimenti sul sistema di interesse, né abbiamo le condizioni che consentono studi quasi sperimentalisti, ci sono ancora molte domande a cui possiamo rispondere con dati puramente osservativi.

Una questione che viene spesso studiata utilizzando dati osservativi è se il decisore abbia utilizzato l'attributo sensibile; questo può essere visto come un vago analogo degli studi di audit. Questo tipo di analisi viene spesso utilizzato nell'analisi giuridica della disparità di trattamento, sebbene esista un dibattito giuridico approfondito e di lunga data su se e quando la considerazione esplicita dell'attributo sensibile sia necessariamente illecita.³¹⁴

Il modo più comune per farlo è utilizzare l'analisi di regressione per vedere se attributi diversi da quelli protetti possono collettivamente "spiegare" le decisioni osservate³¹⁵ (riga 5 nella tabella riepilogativa). In caso contrario, il decisore deve aver utilizzato l'attributo sensibile. Tuttavia, questo è un test fragile. Come discusso nel Capitolo 3, dato un set di dati sufficientemente ricco, l'attributo sensibile può essere ricostruito utilizzando gli altri attributi. Non sorprende che i tentativi di applicare questo test in un contesto legale possano trasformarsi in duelli tra rapporti di esperti, come visto nel caso SFFA vs Harvard discusso nel capitolo 5.

Possiamo ovviamente provare ad andare più in profondità con i dati osservativi e la regressione

analisi. Per illustrare, consideriamo il divario retributivo di genere. Uno studio potrebbe rivelare che esiste un divario tra i sessi nella retribuzione per ora lavorata per posizioni equivalenti in un'azienda. Una confutazione potrebbe affermare che il divario scompare dopo aver controllato il GPA del college e i punteggi di revisione delle prestazioni. Tali studi possono essere visti come test per la parità demografica condizionata (riga 6 nella tabella riassuntiva). Si noti che ciò richiede forti ipotesi sulla forma funzionale della relazione tra le variabili indipendenti e la variabile target.

Può essere difficile dare un senso alle affermazioni concorrenti basate sull'analisi di regressione. Quali variabili dovremmo controllare e perché? Ci sono due modi in cui possiamo porre queste affermazioni osservative su una base più rigorosa. Il primo è utilizzare un quadro causale per rendere le nostre affermazioni più precise. In questo caso, la modellazione causale potrebbe metterci in guardia su domande irrisolte: perché i punteggi di revisione delle prestazioni differiscono in base al genere? Che dire della composizione di genere dei diversi ruoli e livelli di anzianità? L'esplorazione di queste domande potrebbe rivelare pratiche sleali. Naturalmente, in questo caso le domande che abbiamo sollevato sono intuitivamente ovvie, ma altri casi potrebbero essere più complessi.

Il secondo modo per andare più in profondità è applicare la nostra comprensione normativa dell'equità per determinare quali percorsi dal genere al salario siano moralmente problematici. Se il divario retributivo è causato dalle (ben note) differenze di genere nella negoziazione degli aumenti salariali, il datore di lavoro ha la responsabilità morale di mitigarlo? Si tratta ovviamente di una questione normativa e non tecnica.

Test basati sui risultati

Finora in questo capitolo abbiamo presentato molti scenari – screening dei candidati per un posto di lavoro, peer review, udienze sulla libertà condizionale – che hanno una cosa in comune: mentre mirano tutti a prevedere qualche risultato (rendimento lavorativo, qualità della carta, recidiva), il ricercatore non ha accesso ai dati sui risultati reali.

In mancanza di verità di base, l'attenzione si sposta sulle caratteristiche osservabili al momento della decisione, come le qualifiche lavorative. Una persistente fonte di difficoltà in questi contesti è che il ricercatore costruisca due serie di campioni che differiscono solo per l'attributo sensibile e non per nessuna delle caratteristiche rilevanti. Questo è spesso un presupposto non verificabile. Anche in un contesto sperimentale come uno studio di revisione del curriculum, c'è ampio spazio per interpretazioni diverse: i datori di lavoro hanno dedotto la razza dai nomi o dallo stato socioeconomico? E negli studi osservazionali, i risultati potrebbero rivelarsi non validi a causa di fattori confondenti non osservati (come nello studio sui giudici affamati).

Ma se i dati sui risultati sono disponibili, allora possiamo eseguire almeno un test di equità senza bisogno di nessuna delle caratteristiche osservabili (a parte l'attributo sensibile): in particolare, possiamo testare la sufficienza, che richiede che il risultato reale sia condizionatamente indipendente da l'attributo sensibile data la previsione ($Y \mid A|R$).

Ad esempio, nel contesto dei prestiti, se le decisioni della banca soddisfano la sufficienza, allora tra i richiedenti in qualsiasi intervallo ristretto di probabilità di default prevista (R), dovremmo trovare lo stesso tasso di default (Y) per i richiedenti di qualsiasi gruppo (A).

In genere, il decisore (la banca) può verificare la sufficienza, ma un esterno

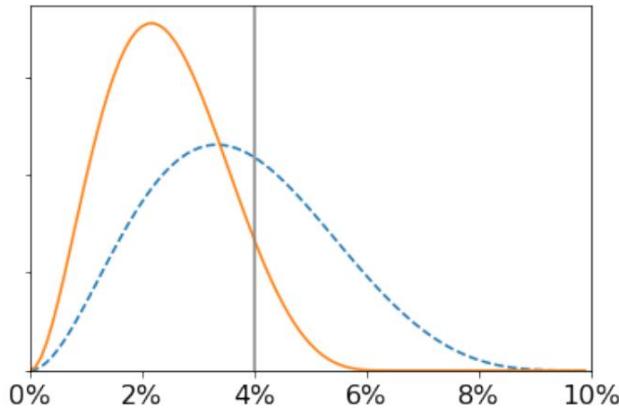


Figura 7.2: Densità di probabilità ipotetica di default del prestito per due gruppi, donne (linea continua) e uomini (linea tratteggiata).

il ricercatore non può, poiché può osservare solo Y e non R (vale a dire, se il prestito è stato approvato o meno). Un ricercatore di questo tipo può testare la parità predittiva piuttosto che la sufficienza. La parità predittiva richiede che il tasso di inadempienza (Y) per i richiedenti classificati favorevolmente ($Y = 1$) di qualsiasi gruppo (A) sia lo stesso. Questo test osservazionale è chiamato test dei risultati (riga 7 nella tabella riepilogativa).

Ecco un argomento allettante basato sul test dei risultati: se un gruppo (ad esempio le donne) che riceve prestiti ha un tasso di insolvenza inferiore rispetto a un altro (uomini), ciò suggerisce che la banca applica un limite più elevato per l'idoneità al prestito per le donne.

In effetti, questo tipo di argomentazione è stata la motivazione originale alla base del test dei risultati. Ma è un errore logico; la sufficienza non implica parità predittiva (o viceversa).

Per capire perché, consideriamo un esperimento mentale che coinvolge il predittore ottimale di Bayes.

Nella figura ipotetica seguente, i richiedenti a sinistra della linea verticale si qualificano per il prestito. Poiché l'area sotto la curva a sinistra della linea è concentrata più a destra per gli uomini che per le donne, gli uomini che ricevono prestiti hanno maggiori probabilità di andare in default rispetto alle donne. Pertanto, il test dei risultati rivelerebbe che la parità predittiva è violata, mentre dalla costruzione è chiaro che la sufficienza è soddisfatta e la banca applica lo stesso criterio a tutti i gruppi.

Questo fenomeno è detto inframarginalità, ovvero la misurazione viene aggregata su campioni lontani dalla soglia decisionale (marginale). Se siamo davvero interessati a testare la sufficienza (equivalentemente, se la banca ha applicato la stessa soglia a tutti i gruppi), piuttosto che la parità predittiva, questo è un problema. Per affrontarlo, possiamo in qualche modo provare a restringere la nostra attenzione ai campioni vicini alla soglia. Ciò non è possibile solo con (Y, A, Y) : senza conoscere R , non sappiamo quali istanze sono vicine alla soglia. Tuttavia, se avessimo accesso anche a qualche insieme di caratteristiche X (che non necessariamente coincidono con l'insieme di caratteristiche X osservate dal decisore), diventa possibile verificare eventuali violazioni della sufficienza. Il test della soglia è un modo per farlo (riga 8 nella tabella riepilogativa). Una descrizione completa va oltre il nostro scopo.³¹⁶ Una limitazione è che richiede un modello della distribuzione congiunta di (X, A, Y) i cui parametri possono essere dedotti dai dati, mentre

il test dei risultati è privo di modelli.

Sebbene abbiamo descritto l'inframarginalità come una limitazione del test di esito, può anche essere vista come un vantaggio. Quando utilizziamo un test marginale, trattiamo la distribuzione delle caratteristiche del candidato come un dato di fatto e perdiamo l'opportunità di chiederci: perché alcuni individui sono così lontani dal margine? Idealmente, possiamo utilizzare l'inferenza causale per rispondere a questa domanda, ma quando i dati a disposizione non lo consentono, i test non marginali potrebbero essere un utile punto di partenza per diagnosticare l'ingiustizia che ha origine "a monte" del decisore. Allo stesso modo, la disparità del tasso di errore, alla quale ci occuperemo ora, sebbene rozza rispetto a test di discriminazione più sofisticati, tenta di catturare alcune delle nostre intuizioni morali sul motivo per cui determinate disparità sono problematiche.

Separazione ed etichette selettive

Ricordiamo che la separazione è definita come $R \not\supseteq A|Y$. A prima vista, sembra che esista un semplice test osservativo analogo al nostro test di sufficienza ($Y \not\supseteq A|R$). Tuttavia, questo non è semplice, nemmeno per il decisore, perché le etichette di risultato possono essere osservate solo per alcuni dei richiedenti (cioè quelli che hanno ricevuto decisioni favorevoli). Il tentativo di testare la separazione utilizzando questo campione soffre di errori di selezione.

Questo è un esempio di quello che viene chiamato il problema delle etichette selettive. Il problema riguarda anche il calcolo della parità di tasso di falsi positivi e falsi negativi, che sono versioni binarie della separazione.

Più in generale, il problema delle etichette selettive è la questione del bias di selezione nella valutazione dei sistemi decisionali dovuto al fatto che lo stesso processo di selezione che desideriamo studiare determina il campione di istanze che vengono osservate. Non è specifico per la questione della separazione dei test o dei tassi di errore: influisce anche sulla misurazione di altri parametri fondamentali come l'accuratezza. Si tratta di una questione seria e spesso trascurata che è stata oggetto di alcuni studi.³¹⁷

Un modo per aggirare questa barriera è che il decisore utilizzi un esperimento in cui alcuni campioni di soggetti decisionali ricevono decisioni positive indipendentemente dalla previsione (riga 9 nella tabella riassuntiva). Tuttavia, tali esperimenti sollevano preoccupazioni etiche e raramente vengono condotti nella pratica. Nell'apprendimento automatico, è necessaria una certa sperimentazione in contesti in cui non esistono dati offline per addestrare il classificatore, che deve invece apprendere e prendere decisioni

simultaneamente.³¹⁸ Uno scenario in cui è semplice testare la separazione è quando la "previsione" è non in realtà una previsione di un evento futuro, ma piuttosto quando l'apprendimento automatico viene utilizzato per automatizzare il giudizio umano, come il rilevamento di molestie nei commenti online. In queste applicazioni è infatti possibile ed importante testare la parità del tasso di errore.

Riepilogo dei test e dei metodi tradizionali

विवरणीयगाली,
विवरणीयनामानुसंधान
विवरणीयनामानुसंधान

		Accesso	Notazioni
1	Cecità	Esposizione R p	Difettare Confusione; विवरणीयानि विवरणीयनामानुसंधान विवरणीयनामानुसंधान
	Arbitrarietà	$\mathbb{R} \setminus Y$	विवरणीयनामानुसंधान विवरणीयनामानुसंधान
	Cecità	\mathbb{R} , \mathbb{R} ,	प्राप्तिशाली विवरणीयनामानुसंधान विवरणीयनामानुसंधान
	विवरणीय विवरणीयनामानुसंधान	\mathbb{X} , विवरणीयनामानुसंधान/े	Inframarginalità प्रतिक्रिया विवरणीयनामानुसंधान
	विवरणीय विवरणीयनामानुसंधान	$\mathbb{R}, := Y$	प्राप्तिशाली विवरणीयनामानुसंधान
	विवरणीय विवरणीयनामानुसंधान	\mathbb{R} ,	विवरणीयनामानुसंधान विवरणीयनामानुसंधान
	विवरणीय विवरणीयनामानुसंधान	\mathbb{R} ,	विवरणीयनामानुसंधान विवरणीयनामानुसंधान

Leggenda:

- := indica un intervento su qualche variabile (ovvero, X := non rappresenta una nuova variabile casuale ma è semplicemente un'annotazione che descrive come X viene utilizzata nel test)
- \dot{y} variazione naturale di alcune variabili sfruttate dal ricercatore • A-exp esposizione di un segnale dell'attributo sensibile al decisore • W una caratteristica considerata irrilevante per la decisione • X un insieme di caratteristiche che potrebbero non coincidere con quelle osservato dal decisore|
- Y un risultato che può o meno essere quello oggetto della previsione|

Discriminazione basata sul gusto e statistica

Abbiamo esaminato diversi metodi per individuare la discriminazione, ma non abbiamo affrontato la questione del perché si verifica la discriminazione. Un metodo utilizzato da molto tempo per cercare di rispondere a questa domanda da una prospettiva economica è quello di classificare la discriminazione come basata sul gusto o statistica. Un discriminatore basato sul gusto è motivato da un'animus irrazionale o da un pregiudizio nei confronti di un gruppo. Di conseguenza, sono disposti a prendere decisioni non ottimali rinunciando all'opportunità di selezionare i candidati di quel gruppo, anche se per farlo incorreranno in una sanzione finanziaria. Questo è il classico modello di discriminazione nel mercato del lavoro introdotto da Gary Becker nel 1957.

Un discriminatore statistico, al contrario, mira a fare previsioni ottimali sulla variabile target utilizzando tutte le informazioni disponibili, compreso l'attributo protetto. Questa teoria è stata sviluppata all'inizio degli anni '70 da Edmund Phelps e Kenneth Arrow, tra gli altri.^{320, 321} Nel modello più semplice di discriminazione statistica, valgono due condizioni: in primo luogo, la distribuzione della variabile target differisce a seconda del gruppo. L'esempio più comune è quello della discriminazione di genere sul posto di lavoro, che coinvolge un datore di lavoro che ritiene che le donne abbiano maggiori probabilità di prendersi delle ferie a causa della gravidanza (con conseguente rendimento lavorativo inferiore). La seconda condizione è che le caratteristiche osservabili non consentano una previsione perfetta della variabile obiettivo, il che nella pratica è sostanzialmente sempre vero. In queste due condizioni, la previsione ottimale differirà da gruppo a gruppo anche quando le caratteristiche rilevanti sono identiche. In questo esempio, il datore di lavoro sarebbe meno propenso ad assumere una donna rispetto a un uomo altrettanto qualificato. C'è una sfumatura qui: da un punto di vista morale diremmo che il datore di lavoro di cui sopra discrimina tutte le candidate donne. Ma secondo la definizione di discriminazione statistica, il datore di lavoro discrimina solo nei confronti delle candidate donne che non avrebbero preso ferie se assunte (e di fatto discrimina a favore delle candidate donne che avrebbero preso ferie se assunte).

Sebbene alcuni autori diano molta importanza alla comprensione della discriminazione basata sulla categorizzazione basata sul gusto rispetto a quella statistica, in questo libro daremo meno importanza a questo aspetto. Diversi motivi motivano la nostra scelta. In primo luogo, poiché siamo interessati a trarre lezioni per i sistemi decisionali statistici, la distinzione non è così utile: tali sistemi non mostreranno una discriminazione basata sul gusto a meno che non vi siano pregiudizi

è esplicitamente programmato in essi (anche se questa è certamente una possibilità, non è una delle principali preoccupazioni di questo libro).

In secondo luogo, esistono difficoltà pratiche nel distinguere tra discriminazione basata sul gusto e discriminazione statistica. Spesso, quello che potrebbe sembrare un “gusto” per la discriminazione è semplicemente il risultato di una comprensione imperfetta delle informazioni e delle convinzioni del decisore. Ad esempio, a prima vista i risultati dello studio sulla contrattazione automobilistica potrebbero sembrare un chiaro caso di discriminazione basata sul gusto. Ma forse il rivenditore sa che clienti diversi hanno un accesso diverso alle offerte concorrenti e quindi hanno una diversa disponibilità a pagare per lo stesso articolo. Quindi, il dealer utilizza la razza come proxy per questo importo (correttamente o meno). In effetti, il documento fornisce prove provvisorie verso questa interpretazione. È possibile anche il contrario: se il ricercatore non conosce l’insieme completo delle caratteristiche osservate dal decisore, la discriminazione basata sul gusto potrebbe essere erroneamente definita discriminazione statistica.

In terzo luogo, molte delle questioni di equità che ci interessano, come la discriminazione strutturale, non corrispondono a nessuno di questi criteri (poiché considerano solo cause relativamente vicine al punto decisionale). Discuteremo della discriminazione strutturale nel capitolo 8.

Infine, vale anche la pena notare che pensare alla discriminazione in termini di dicotomia tra basata sul gusto e statistica è associato alla posizione politica secondo cui gli interventi sull’equità non sono necessari. In quest’ottica, le aziende che praticano la discriminazione basata sul gusto falliranno. Per quanto riguarda la discriminazione statistica, si ritiene che sia giustificato o inutile vietarla perché le aziende troveranno soluzioni alternative. Ad esempio, le leggi che vietano ai datori di lavoro di chiedere informazioni sui precedenti penali dei candidati hanno portato i datori di lavoro a utilizzare la razza come indicatore.³²²

Naturalmente, questo non è necessariamente un motivo per evitare di discutere di discriminazione basata sul gusto e statistica, poiché la posizione politica non deriva in alcun modo dalle definizioni e dai modelli tecnici stessi; è solo un avvertimento rilevante per il lettore che potrebbe incontrare questi dubbi argomenti in altre fonti.

Anche se sottolineiamo questa distinzione, riteniamo fondamentale studiare le fonti e i meccanismi della discriminazione. Questo ci aiuta a progettare interventi efficaci e ben mirati. Ad esempio, diversi studi (incluso lo studio sulla contrattazione automobilistica) verificano se la fonte della discriminazione risiede nel proprietario, nei dipendenti o clienti.

Un esempio di studio che può essere difficile da interpretare senza comprenderne il meccanismo è uno studio di audit basato sul curriculum del 2015 che ha rivelato una preferenza 2:1 tra i docenti per le posizioni di ruolo STEM.³²³ Considerate la gamma di possibili spiegazioni: animosità contro uomini; il desiderio di compensare gli svantaggi passati subiti dalle donne nei campi STEM; una preferenza per una facoltà più diversificata (assumendo che le facoltà in questione siano attualmente a predominanza maschile); una risposta agli incentivi finanziari per la diversificazione spesso forniti dalle università ai dipartimenti STEM; e l’ipotesi da parte dei decisori che, a causa della precedente discriminazione, una candidata donna con un CV equivalente a un candidato uomo abbia maggiori capacità intrinseche. Si noti che se questo presupposto è corretto, allora la preferenza per le candidate donne massimizza l’accuratezza (come predittore del successo professionale). È richiesto anche da alcuni criteri di equità, come l’equità controllattuale.

Per riassumere, piuttosto che un approccio unico per tutti per comprendere meccanismi come la discriminazione basata sul gusto o quella statistica, più utile è un approccio sfumato e specifico per dominio in cui formuliamo ipotesi in parte studiando i processi decisionali e organizzazioni, soprattutto in termini qualitativi. Passiamo ora a quegli studi.

Studi sui processi decisionali e sulle organizzazioni

Un modo per studiare i processi decisionali è attraverso sondaggi tra decisori o organizzazioni. Talvolta tali studi rivelano una palese discriminazione, come ad esempio forti preferenze razziali da parte dei datori di lavoro.³²⁴ Nel corso dei decenni, tuttavia, tali atteggiamenti palesi sono diventati meno comuni, o almeno hanno meno probabilità di essere espressi.³²⁵ La discriminazione tende ad operare in modo più sottile, indiretto, e modi nascosti.

Gli studi etnografici eccellono nell'aiutarci a comprendere la discriminazione nascosta. L'etnografia è uno dei principali metodi di ricerca nelle scienze sociali e si basa sull'idea che il ricercatore sia inserito tra i soggetti di ricerca per un lungo periodo di tempo mentre svolgono le loro attività quotidiane. Si tratta di un insieme di metodi qualitativi complementari e simbiotici con quelli quantitativi. L'etnografia ci consente di porre domande che sono più profonde di quanto consentano i metodi quantitativi e di produrre resoconti della cultura riccamente dettagliati. Aiuta anche a formulare ipotesi che possono essere testate quantitativamente.

Un buon esempio è il libro *Pedigree* di Lauren Rivera che esamina le pratiche di assunzione in una serie di studi legali, bancari e di consulenza d'élite.³²⁶ Queste aziende insieme costituiscono la maggior parte dei lavori entry-level più pagati e più desiderabili per i laureati. L'autore ha utilizzato due metodi di ricerca etnografica standard.

La prima è una serie di 120 interviste in cui si è presentata come una studentessa laureata interessata a opportunità di stage. Il secondo metodo è chiamato osservazione partecipante: ha lavorato per 9 mesi in una posizione non retribuita nel settore delle risorse umane presso una delle aziende, dopo aver ottenuto il consenso per utilizzare le sue osservazioni per la ricerca. Ci sono diversi vantaggi nel fatto che il ricercatore diventi partecipe della cultura: fornisce un maggiore livello di accesso, consente al ricercatore di porre domande più sfumate e rende più probabile che i soggetti della ricerca si comportino come farebbero quando non vengono osservati.

Molti spunti del libro sono rilevanti per noi. Innanzitutto, il processo di assunzione prevede circa nove fasi, tra cui sensibilizzazione, eventi di reclutamento, screening, molteplici cicli di interviste e deliberazioni ed eventi di "vendita". Ciò evidenzia il motivo per cui qualsiasi studio quantitativo che si concentra su una singola parte del processo (ad esempio, la valutazione del curriculum) ha una portata limitata. In secondo luogo, il processo ha poca somiglianza con l'ideale di prevedere la prestazione lavorativa sulla base di un insieme standardizzato di attributi, anche se rumorosi, che abbiamo descritto nel capitolo 1. Gli intervistatori prestano una sorprendente quantità di attenzione agli attributi che dovrebbero essere irrilevanti o minimamente rilevanti, come attività del tempo libero, ma che servono invece come indicatori di classe. I candidati provenienti da contesti privilegiati hanno maggiori probabilità di essere visti favorevolmente, sia perché possono dedicare più tempo a tali attività, sia perché hanno informazioni privilegiate

consapevolezza che questi attributi apparentemente irrilevanti contano nel reclutamento. I segnali che le aziende utilizzano come predittori della performance lavorativa, come l'ammissione alle università d'élite – il pedigree nel titolo del libro – sono anche altamente correlati con lo status socioeconomico. Gli autori sostengono che queste pratiche di assunzione aiutano a spiegare perché lo status di élite si perpetua nella società lungo linee ereditarie. A nostro avviso, l'uso attento dei metodi statistici nelle assunzioni, nonostante i loro limiti, può mitigare le forti preferenze basate sulla classe sociale esposte nel libro.

Un altro libro, *Inside Graduate Admissions* di Julie Posselt, si concentra sull'istruzione piuttosto che sul mercato del lavoro.³²⁷ È il risultato delle osservazioni dell'autore sul processo decisionale dei comitati di ammissione dei laureati in nove discipline accademiche nell'arco di due anni. Un tema sorprendente che pervade questo libro è la tensione tra il processo decisionale formalizzato e quello olistico. Ad esempio, i comitati fanno probabilmente eccessivo affidamento sui punteggi GRE nonostante affermino di considerare limitato il loro potere predittivo. A quanto pare, uno dei motivi della preferenza per i punteggi GRE e altri criteri quantitativi è che evitano le difficoltà di interpretazione soggettiva associate a segnali come le lettere di referenza. Ciò è considerato prezioso perché riduce al minimo le tensioni tra i membri della facoltà nel processo di ammissione. D'altro canto, i decisori sono implicitamente consapevoli (e talvolta lo esprimono esplicitamente) che se i criteri di ammissione fossero troppo formali, allora alcuni gruppi di candidati – in particolare quelli provenienti dalla Cina – avrebbero successo a un ritmo molto maggiore, e questo è considerato indesiderabile. .. Ciò motiva un insieme di criteri più olistici, che spesso includono fattori peculiari come l'hobby del candidato considerato "interessante" da un membro della facoltà. L'autore sostiene che i comitati di ammissione utilizzano una serie di criteri apparentemente neutri, caratterizzati da un'assenza quasi totale di discussione esplicita e sostanziale sulla razza, il genere o lo status socioeconomico dei candidati, ma che comunque perpetua le disuguaglianze. Ad esempio, c'è una certa riluttanza ad assumere studenti provenienti da contesti sottorappresentati i cui profili suggeriscono che trarrebbero beneficio da un tutoraggio più intenso.

Con questo si conclude la prima parte del capitolo. Passiamo ora ai sistemi algoritmici. Il background che abbiamo costruito finora si rivelerà utile. In effetti, i tradizionali test di discriminazione sono altrettanto applicabili ai sistemi algoritmici. Ma incontreremo anche molte questioni nuove.

Parte 2: Testare la discriminazione nei sistemi algoritmici

Un primo esempio di discriminazione in un sistema algoritmico risale agli anni '50. Negli Stati Uniti, i richiedenti per i programmi di residenza medica forniscono un elenco classificato dei loro programmi ospedalieri preferiti a un sistema centralizzato e anche gli ospedali classificano i candidati. Un algoritmo di corrispondenza prende queste preferenze come input e produce un'assegnazione dei candidati agli ospedali che ottimizza la reciproca desiderabilità.³

³In particolare, soddisfa il requisito secondo cui se il richiedente A non è abbinato all'ospedale H, allora A è abbinato a un ospedale che ha classificato più in alto di H, oppure H è abbinato a un insieme di richiedenti per i quali ha tutti classificato più alto di A.

Le prime versioni del sistema discriminavano le coppie che desideravano rimanere geograficamente vicine, perché le coppie non potevano esprimere accuratamente le loro preferenze comuni: ad esempio, ciascun partner poteva preferire un ospedale rispetto a tutti gli altri, ma solo se anche l'altro partner corrispondeva allo stesso ospedale. 328, 43 Si tratta di una nozione di discriminazione non comparativa: il sistema fa un'ingiustizia nei confronti di un richiedente (o di una coppia) quando non consente loro di esprimere le proprie preferenze, indipendentemente da come vengono trattati gli altri richiedenti. Si noti che nessuno dei test di equità di cui abbiamo discusso è in grado di rilevare questo caso di discriminazione, poiché nasce a causa delle dipendenze tra coppie di unità, che non è qualcosa che abbiamo modellato.

C'è stato un tentativo grossolano nel sistema di corrispondenza delle residenze di catturare le preferenze comuni, implicando la designazione di un partner in ciascuna coppia come "membro principale"; l'algoritmo abbinerebbe il membro principale senza vincoli e quindi abbinerebbe l'altro membro a un ospedale vicino, se possibile. Date le norme di genere prevalenti a quel tempo, è probabile che questo metodo abbia avuto un ulteriore impatto discriminatorio sulle donne nelle coppie eterosessuali.

Nonostante questi primi esempi, è a partire dagli anni 2010 che testare l'ingiustizia nei sistemi algoritmici del mondo reale è diventato una preoccupazione urgente e un'area di ricerca distinta. Questo lavoro ha molto in comune con la ricerca sulle scienze sociali che abbiamo esaminato, ma gli obiettivi della ricerca si sono notevolmente ampliati. Nel resto di questo capitolo esamineremo e tenteremo di sistematizzare i metodi di ricerca in diverse aree del processo decisionale algoritmico: varie applicazioni dell'elaborazione del linguaggio naturale e della visione artificiale; piattaforme di targeting degli annunci; strumenti di ricerca e recupero di informazioni; e mercati online (ride hailing, affitti per vacanze, ecc.). Gran parte di questa ricerca si è concentrata sull'attirare l'attenzione sugli effetti discriminatori di strumenti e piattaforme specifici e ampiamente utilizzati in momenti specifici nel tempo. Sebbene si tratti di un obiettivo prezioso, nella nostra recensione mireremo a evidenziare temi più ampi e generalizzabili. Chiuderemo il capitolo identificando principi e metodi comuni alla base di questo corpo di ricerca.

Considerazioni sull'equità nelle applicazioni dell'elaborazione del linguaggio naturale

Uno dei compiti più centrali nella PNL è l'identificazione della lingua: determinare la lingua in cui è scritto un determinato testo. È un precursore praticamente di qualsiasi altra operazione di PNL sul testo, come la traduzione nella lingua preferita dell'utente sulle piattaforme di social media. È considerato un problema più o meno risolto, con modelli relativamente semplici basati su n grammi di caratteri che raggiungono un'elevata precisione rispetto ai parametri di riferimento standard, anche per testi brevi di poche parole.

Tuttavia, uno studio del 2016 ha mostrato che uno strumento ampiamente utilizzato, langid.py, che incorpora un modello pre-addestrato, aveva sostanzialmente più falsi negativi per i tweet scritti in inglese afro-americano (AAE) rispetto a quelli scritti nelle forme dialettali più comuni: Il 13,2% dei tweet AAE sono stati classificati come non inglese rispetto al 7,6% dei tweet inglesi "allineati ai bianchi". AAE è un insieme di dialetti inglesi comunemente parlati dai neri negli Stati Uniti (ovviamente li

non implica che tutti i neri negli Stati Uniti parlino principalmente AAE o addirittura lo parlino affatto⁴. La costruzione degli stessi corpora AAE e White-aligned da parte degli autori ha coinvolto l'apprendimento automatico e la convalida basata sull'esperienza linguistica; rinviamo la trattazione completa al capitolo Misurazione. La disparità osservata nel tasso di errore è probabilmente un classico caso di sottorappresentazione nei dati di addestramento.

A differenza degli studi di audit sulle vendite di automobili o sui mercati del lavoro discussi in precedenza, in questo caso non è necessario (o giustificabile) controllare alcuna caratteristica dei testi, come il livello di formalità. Sebbene sia certamente possibile "spiegare" tassi di errore disparati sulla base di tali caratteristiche, ciò è irrilevante per le questioni di interesse in questo contesto, ad esempio se gli strumenti di PNL funzioneranno meno bene per un gruppo di utenti rispetto a un altro.

Gli strumenti della PNL spaziano nella loro applicazione da ausili per l'interazione online a componenti di decisioni con importanti conseguenze sulla carriera. In particolare, la PNL viene utilizzata negli strumenti predittivi per lo screening del curriculum nel processo di assunzione. Esistono prove del potenziale impatto discriminatorio di tali strumenti, sia da parte degli stessi datori di lavoro³³⁰ che da parte dei candidati³³¹, ma si limitano ad aneddoti. Esistono anche prove provenienti dagli esperimenti di laboratorio sul compito di prevedere l'occupazione dalle biografie

online.³³² Esaminiamo brevemente altri risultati. I software di valutazione automatizzata dei saggi tendono ad assegnare sistematicamente punteggi più bassi ad alcuni gruppi demografici³³³ rispetto ai valutatori umani, i cui punteggi potrebbero essere essi stessi discriminatori.³³⁴ Secondo uno studio di laboratorio, i modelli di rilevamento dell'incitamento all'odio utilizzano marcatori dialettali come predittori di tossicità,³³⁵ con conseguente discriminazione contro i parlanti di minoranza. Molti strumenti di analisi del sentimento assegnano sistematicamente punteggi diversi al testo in base ai nomi delle persone menzionate nel testo in base alla razza o al genere.³³⁶ I sistemi di sintesi vocale funzionano peggio per chi parla con determinati accenti.³³⁷ In tutti questi casi, l'autore o chi parla il testo è potenzialmente danneggiato. In altri sistemi di PNL, cioè quelli che coinvolgono la generazione o la traduzione del linguaggio naturale, esiste un diverso tipo di preoccupazione per l'equità, vale a dire la generazione di testo che riflette pregiudizi culturali con conseguente danno rappresentazionale per un gruppo di persone.³³⁸ La tabella seguente riporta una sintesi di questi risultati.

Esiste una linea di ricerca sugli stereotipi culturali riflessi negli incorporamenti di parole. Gli incorporamenti di parole sono rappresentazioni di unità linguistiche; non corrispondono ad alcun compito linguistico o decisionale. In quanto tale, in mancanza di qualsiasi nozione di verità fondamentale o di danni alle persone, non ha senso porre domande sull'equità sugli incorporamenti di parole senza riferimento a specifici compiti a valle in cui potrebbero essere utilizzati. Più in generale, non ha senso attribuire l'equità a un attributo dei modelli in contrapposizione ad azioni, risultati o processi decisionali.

⁴Per un trattato sull'AAE, vedere.³²⁹ Lo studio linguistico dell'AAE evidenzia la complessità e la coerenza interna della sua grammatica, del vocabolario e di altre caratteristiche distintive, e confuta la base di visioni pregiudiziali dell'AAE come inferiore all'inglese standard.

Prevedere.	Prevedere.	Dannno

Disparità demografiche e applicazioni discutibili della visione artificiale

Come la PNL, la tecnologia della visione artificiale ha fatto grandi progressi negli anni 2010 grazie alla disponibilità di corpora di formazione su larga scala e ai miglioramenti nell'hardware per l'addestramento delle reti neurali. Oggi, nei prodotti commerciali vengono utilizzati molti tipi di classificatori per analizzare immagini e video di persone. Non sorprende che spesso mostrino disparità nelle prestazioni basate sul genere, sulla razza, sul tono della pelle e su altri attributi, oltre a problemi etici più profondi.

Un'importante dimostrazione della disparità nel tasso di errore viene dall'analisi di Buolamwini e Gebril di tre strumenti commerciali progettati per classificare il genere di una persona come femminile o maschile sulla base di un'immagine, sviluppati rispettivamente da Microsoft, IBM e Face++.³³⁹ Lo studio ha rilevato che tutti e tre i classificatori ottengono risultati migliori sui volti maschili rispetto ai volti femminili (differenza tra 8,1% e 20,6% nel tasso di errore). Inoltre, tutti hanno risultati migliori sui volti più chiari rispetto ai volti più scuri (differenza dell'11,8% – 19,2% nel tasso di errore) e peggiori sui volti femminili più scuri (tasso di errore del 20,8% – 34,7%). Infine, poiché tutti i classificatori trattano il genere come binario, il tasso di errore per le persone di genere non binario può essere considerato pari al 100%.

Se trattiamo la variabile target del classificatore come il sesso e l'attributo sensibile come il tono della pelle, possiamo scomporre le disparità osservate in due questioni separate: in primo luogo, i volti femminili sono classificati come maschili più spesso di quanto i volti maschili siano classificati come femminili. Questo problema può essere risolto in modo relativamente semplice ricalibrando la soglia di classificazione senza modificare il processo di formazione. Il secondo e più profondo problema è che i volti più scuri vengono classificati erroneamente più spesso rispetto ai volti più chiari.

Gli strumenti di classificazione delle immagini hanno trovato particolarmente difficile raggiungere l'equità geografica a causa della distorsione dei set di dati di addestramento. Uno studio del 2019 ha valutato cinque popolari servizi di riconoscimento di oggetti su immagini di oggetti domestici provenienti da 54 paesi.³⁴⁰ Ha riscontrato significative disparità di accuratezza tra paesi, con immagini provenienti da paesi a basso reddito classificate in modo meno accurato. Gli autori sottolineano che gli oggetti domestici come il detersivo per i piatti o i contenitori per le spezie tendono ad apparire molto diversi nei diversi paesi. Questi problemi sono esacerbati quando le immagini delle persone vengono classificate. Un'analisi del 2017 ha rilevato che i modelli addestrati su ImageNet e Open Images, due importanti set di dati per il riconoscimento degli oggetti, hanno ottenuto risultati notevolmente peggiori nel riconoscere le immagini degli sposi provenienti da paesi come il Pakistan e l'India rispetto a quelli dei paesi nordamericani ed europei (i primi erano spesso classificati come cotta di maglia, un tipo di armatura).³⁴¹

Molti altri tipi di ingiustizia sono noti attraverso prove aneddotiche nei sistemi di classificazione delle immagini e di riconoscimento facciale. È noto che almeno due diversi sistemi di classificazione delle immagini hanno applicato etichette umilianti e offensive alle foto di persone.^{342, 343} È stato riportato aneddoticamente che i sistemi di riconoscimento facciale mostrino l'effetto razziale in cui sono più propensi a confondere i volti di due persone che appartengono a un gruppo razziale sottorappresentato nei dati di addestramento.³⁴⁴ Questa possibilità è stata mostrata in un semplice modello lineare di riconoscimento facciale già nel 1991.

³⁴⁵ Molti prodotti commerciali hanno avuto difficoltà nel rilevare i volti di persone dalla pelle più scura.^{346, 347} Risultati simili sono noti da studi di laboratorio su soggetti pubblici.

modelli di rilevamento oggetti disponibili.³⁴⁸

Più in generale, le tecniche di visione artificiale sembrano essere particolarmente inclini a essere utilizzate in modi che sono fondamentalmente eticamente discutibili, indipendentemente dalla loro accuratezza. Considera la classificazione di genere: mentre Microsoft, IBM e Face++ hanno lavorato per mitigare le disparità di accuratezza discusse sopra, una domanda più importante è in primo luogo perché costruire uno strumento di classificazione di genere. L'applicazione di gran lunga più comune sembra essere la visualizzazione di annunci pubblicitari mirati basati sul genere dedotto (e molte altre caratteristiche dedotte, tra cui età, razza e umore attuale) negli spazi pubblici, come cartelloni pubblicitari, negozi o schermi sui sedili posteriori dei taxi. Non ricapitoliamo qui le obiezioni alla pubblicità mirata, ma si tratta di un argomento ampiamente discusso e la pratica è fortemente osteggiata dal pubblico, almeno negli Stati Uniti.³⁴⁹

La tecnologia di visione artificiale, moralmente dubbia, va ben oltre questo esempio e include app che "abbelliscono" le immagini dei volti degli utenti, ovvero le modificano per conformarsi meglio alle nozioni tradizionali di attrattiva; riconoscimento delle emozioni, che è stato ritenuto una pseudoscienza; e l'analisi delle riprese video per individuare segnali come il linguaggio del corpo per lo screening dei candidati al lavoro.³⁵⁰

Sistemi di ricerca e raccomandazione: tre tipologie di danni

I motori di ricerca, le piattaforme di social media e i sistemi di raccomandazione hanno obiettivi e algoritmi sottostanti diversi, ma hanno molte cose in comune dal punto di vista dell'equità. Non sono sistemi decisionali e non forniscono né negano opportunità alle persone, almeno non direttamente. Invece, ci sono (almeno) tre tipi di disparità e di conseguenti danni che possono verificarsi in questi sistemi.

In primo luogo, potrebbero soddisfare le esigenze informative di alcuni consumatori (ricercatori o utenti) meglio di altri. In secondo luogo, possono creare disuguaglianze tra i produttori (creatori di contenuti) privilegiando determinati contenuti rispetto ad altri. In terzo luogo, possono creare danni rappresentazionali amplificando e perpetuando gli stereotipi culturali. Esistono numerose altre preoccupazioni etiche relative alle piattaforme di informazione, come il potenziale di contribuire alla polarizzazione politica della società. Limiteremo però la nostra attenzione ai danni che possono essere considerati forme di discriminazione.

Ingiustizia nei confronti dei consumatori. Un esempio di ingiustizia nei confronti dei consumatori viene da uno studio sui sistemi di raccomandazione di filtraggio collaborativo che utilizzavano metodi teorici e di simulazione (piuttosto che uno studio sul campo di un sistema implementato).³⁵¹ Il filtraggio collaborativo è un approccio alle raccomandazioni basato sull'esplicito o feedback impliciti (ad esempio valutazioni e consumi, rispettivamente) forniti da altri utenti del sistema. L'intuizione che c'è dietro si vede nel fatto che "gli utenti a cui è piaciuto questo articolo hanno apprezzato anche" . . . " è presente su molti servizi. Lo studio ha rilevato che tali sistemi possono sottoperformare per i gruppi minoritari, nel senso che sono peggiori nel consigliare i contenuti che quegli utenti vorrebbero. Una ragione correlata ma distinta per la sottoperformance si verifica quando gli utenti di un gruppo sono meno osservabili, ad esempio, hanno meno probabilità di fornire valutazioni. Il presupposto di fondo è che gruppi diversi abbiano preferenze diverse, in modo che ciò che il sistema apprende su un gruppo non si generalizzi

altri gruppi.

In generale, questo tipo di ingiustizia è difficile da studiare nei sistemi reali (non solo da ricercatori esterni ma anche dagli stessi operatori di sistema). La difficoltà principale è misurare accuratamente la variabile target. Il costrutto target rilevante dal punto di vista dell'equità è la soddisfazione degli utenti con i risultati o il modo in cui i risultati hanno soddisfatto le esigenze degli utenti. Metriche come clic e valutazioni fungono da indicatori grezzi per il target e sono essi stessi soggetti a bias di misurazione demografica. Le aziende investono risorse significative in test A/B o altri metodi sperimentali per ottimizzare i sistemi di ricerca e raccomandazione e spesso misurano anche le differenze demografiche. Ma, per ribadirlo, tali test enfatizzano quasi sempre i parametri di interesse per l'azienda piuttosto che i benefici o i guadagni per l'utente.

Un raro tentativo di trascendere questa limitazione viene da uno studio di audit (interno) del motore di ricerca Bing condotto da Merhotra et al.³⁵² Gli autori hanno ideato metodi per distinguere la soddisfazione degli utenti da altre variazioni demografiche specifiche controllando gli effetti dei fattori demografici sulle metriche comportamentali. Lo hanno combinato con un metodo per dedurre direttamente le differenze latenti invece di stimare la soddisfazione degli utenti per ciascun gruppo demografico e quindi confrontare queste stime. Questo metodo deduce quale impressione, tra una coppia di impressioni selezionate casualmente, ha portato a una maggiore soddisfazione dell'utente. Lo hanno fatto utilizzando indicatori di soddisfazione come il tasso di riformulazione. La riformulazione di una query di ricerca è un forte indicatore di insoddisfazione per i risultati. Sulla base di questi metodi, non hanno riscontrato differenze di genere nella soddisfazione, ma lievi differenze di età.

Ingiustizia verso i produttori. Nel 2019, un gruppo di creatori di contenuti ha citato in giudizio YouTube sostenendo che gli algoritmi di YouTube e i moderatori umani hanno soppresso la portata dei video incentrati su LGBT e la capacità di guadagnare entrate pubblicitarie da essi. Questo è un tipo di problema distinto da quello discusso sopra, poiché l'affermazione riguarda un danno ai produttori piuttosto che ai consumatori (sebbene, ovviamente, anche gli spettatori di YouTube interessati ai contenuti LGBT siano presumibilmente danneggiati). Ci sono molte altre accuse e controversie in corso che rientrano in questa categoria: pregiudizi partigiani nei risultati di ricerca e nelle piattaforme di social media, motori di ricerca che privilegiano i risultati delle proprie proprietà rispetto ai concorrenti, verifica dei fatti degli annunci politici online e inadeguatezza (o, al contrario, eccessivamente aggressivo) controllo di presunte violazioni del copyright. È difficile discutere e affrontare in modo significativo questi problemi attraverso la lente dell'equità e della discriminazione piuttosto che una prospettiva più ampia di potere e responsabilità. La questione centrale è che quando le piattaforme di informazione hanno il controllo sul discorso pubblico, diventano arbitri dei conflitti tra interessi e punti di vista concorrenti. Da un punto di vista legale, queste questioni rientrano principalmente nella legge antitrust e nella regolamentazione delle telecomunicazioni piuttosto che nella legge antidiscriminativa.

Danni rappresentazionali. Il libro *Algorithms of Oppression* ha attirato l'attenzione sui modi in cui i motori di ricerca rafforzano dannosi stereotipi razziali, di genere e intersezionali.⁴⁸ Sono stati condotti anche studi quantitativi su alcuni aspetti di questi danni. In linea con il nostro focus quantitativo, discutiamo di uno studio che ha misurato quanto bene la distorsione di genere nei risultati di ricerca di immagini di Google per 45 occupazioni (autore, barista, operaio edile...) corrispondesse alla distorsione di genere nel mondo reale delle rispettive occupazioni.³⁷ Questo può essere visto come un test di calibrazione: le istanze lo sono

occupazioni e la percentuale di donne nei risultati di ricerca è vista come un preditore della percentuale di donne nell'occupazione nel mondo reale. Lo studio ha rilevato prove deboli di esagerazione degli stereotipi, ovvero gli squilibri nelle statistiche occupazionali sono esagerati nei risultati di ricerca di immagini. Tuttavia, le deviazioni erano minori.

Consideriamo un esperimento mentale: supponiamo che lo studio non abbia trovato prove di errata calibrazione. Il sistema risultante è giusto? Sarebbe semplicistico rispondere affermativamente per almeno due ragioni. Innanzitutto, lo studio ha testato la calibrazione tra i risultati della ricerca di immagini e le statistiche occupazionali negli Stati Uniti. Gli stereotipi di genere delle occupazioni e le statistiche occupazionali differiscono sostanzialmente tra paesi e culture. In secondo luogo, riflettere accuratamente le statistiche del mondo reale può comunque costituire un danno rappresentazionale quando tali statistiche sono distorte e riflettono esse stesse una storia di pregiudizi. Un sistema di questo tipo contribuisce alla mancanza di modelli visibili per i gruppi sottorappresentati. Fino a che punto le piattaforme di informazione debbano assumersi la responsabilità di ridurre al minimo questi squilibri e quali tipi di interventi siano giustificati rimangono oggetto di dibattito.

Comprendere l'ingiustizia nel targeting degli annunci

Gli annunci pubblicitari sono stati a lungo presi di mira in modi relativamente rozzi. Ad esempio, una rivista sanitaria potrebbe pubblicare annunci di prodotti di bellezza, sfruttando una correlazione grossolana. A differenza dei metodi precedenti, il targeting online offre numerosi vantaggi chiave agli inserzionisti: raccolta di dati granulari sugli individui, capacità di raggiungere un pubblico di nicchia (in teoria, la dimensione del pubblico può essere una, poiché il contenuto dell'annuncio può essere generato in modo programmatico e personalizzato con attributi utente come input) e la capacità di misurare la conversione (la conversione avviene quando qualcuno che visualizza l'annuncio fa clic su di esso e poi esegue un'altra azione, ad esempio un acquisto). Ad oggi, il targeting degli annunci è stata una delle applicazioni dell'apprendimento automatico di maggior impatto commerciale.

La complessità del moderno targeting degli annunci si traduce in molte strade per disparità nei dati demografici delle visualizzazioni degli annunci, che studieremo. Ma non è ovvio come collegare queste disparità all'equità. Dopotutto, molti tipi di targeting demografico, come gli annunci di abbigliamento in base al sesso, sono considerati innocui.

Esistono due quadri per comprendere i potenziali danni derivanti dal targeting degli annunci. Il primo quadro vede gli annunci come opportunità di sblocco per i loro destinatari, perché forniscono informazioni che lo spettatore potrebbe non avere. Questo è il motivo per cui indirizzare annunci di lavoro o alloggi basati su categorie protette può essere ingiusto e illegale. Gli ambiti in cui gli attacchi mirati sono legalmente vietati corrispondono ampiamente a quelli che hanno un impatto sui diritti civili e riflettono le complesse storie di discriminazione in tali ambiti, come discusso nel capitolo 6.

Il secondo quadro considera gli annunci come strumenti di persuasione piuttosto che di diffusione delle informazioni. In questo contesto, i danni derivano dal fatto che gli annunci sono manipolativi – cioè esercitano un'influenza nascosta invece di fare appelli esplicativi – o sfruttano gli stereotipi.³⁵³ Gli utenti vengono danneggiati dall'essere presi di mira con annunci che forniscono loro un'utilità negativa, al contrario del primo quadro, in cui il danno deriva dalla perdita di annunci con utilità positiva. I due framework non lo fanno necessariamente

contraddirsi tra loro. Piuttosto, singoli annunci o campagne pubblicitarie possono essere visti come principalmente informativi o principalmente persuasivi e, di conseguenza, l'uno o l'altro quadro potrebbe essere appropriato per

l'analisi.⁵ Esiste una vasta letteratura su come la razza e il genere vengono rappresentati nelle pubblicità; riteniamo che questa letteratura rientri nel quadro della persuasione.³⁵⁵ Tuttavia, questa linea di indagine deve ancora rivolgere la sua attenzione alla pubblicità mirata online, che ha il potenziale per accentuare i danni della manipolazione e degli stereotipi prendendo di mira persone e gruppi specifici. Pertanto, la ricerca empirica che metteremo in evidenza rientra nel quadro informativo.

Esistono circa tre meccanismi attraverso i quali lo stesso annuncio mirato può raggiungere un gruppo più spesso di un altro. Il più ovvio è l'uso di criteri di targeting esplicativi da parte degli inserzionisti: l'attributo sensibile stesso o un proxy per esso (come il codice postale come proxy per la razza). Ad esempio, Facebook consente migliaia di categorie di targeting, comprese le categorie costruite automaticamente dal sistema in base alle descrizioni testuali in formato libero dei loro interessi da parte degli utenti. Le indagini di ProPublica hanno scoperto che queste categorie includevano "odiatori di ebrei" e molti altri termini antisemiti.³⁵⁶ L'azienda ha avuto difficoltà a eliminare anche i proxy diretti per categorie sensibili, con conseguenti ripetute denunce.

Il secondo meccanismo che produce disparità è l'ottimizzazione del tasso di clic (o un'altra misura di efficacia), che è uno degli obiettivi principali del targeting algoritmico. A differenza della prima categoria, questa non richiede un intento esplicito da parte dell'inserzionista o della piattaforma. Il sistema algoritmico può prevedere la probabilità di un utente di interagire con un annuncio in base al suo comportamento passato, ai suoi interessi espressi e ad altri fattori (inclusi, potenzialmente, attributi sensibili esplicitamente espressi).

Il terzo meccanismo riguarda gli effetti di mercato: offrire un annuncio a utenti diversi può costare all'inserzionista importi diversi. Ad esempio, alcuni ricercatori hanno osservato che fare pubblicità alle donne costa di più rispetto agli uomini e hanno ipotizzato che ciò sia dovuto al fatto che le donne cliccavano più spesso sugli annunci, il che porta a una maggiore misura di efficacia.^{274, 357} Pertanto, se l'inserzionista specifica semplicemente un budget totale e lascia la consegna alla piattaforma (che è una pratica comune), quindi la composizione del pubblico varierà a seconda del budget: budget inferiori comporteranno una sovrarappresentanza del gruppo meno costoso.

In termini di metodi per rilevare queste disparità, ricercatori e giornalisti hanno utilizzato sostanzialmente due approcci: interagire con il sistema sia come utente che come inserzionista. Datta, Tschantz e Datta hanno creato utenti simulati che avevano l'attributo "sesso" nella pagina Impostazioni annunci di Google impostato su femmina o maschio e hanno scoperto che Google mostrava agli utenti simulati di sesso maschile annunci di una certa agenzia di coaching professionale che prometteva alti stipendi più frequentemente rispetto agli utenti di sesso maschile simulati. utenti di sesso femminile simulati.³⁵⁸ Anche se questo tipo di studio stabilisce che gli annunci di lavoro attraverso il sistema di annunci di Google non tengono conto del genere (come espresso nella pagina delle impostazioni degli annunci), non può scoprire il meccanismo, vale a dire, distinguere tra targeting esplicito da parte dell'inserzionista e piattaforma effetti di varia natura.

Interagire con le piattaforme pubblicitarie come inserzionista si è rivelato più fruttuoso

⁵L'analisi economica della pubblicità comprende una terza categoria, complementare, alla quale è correlata categoria persuasiva o manipolativa.³⁵⁴

approccio finora, in particolare per analizzare il sistema pubblicitario di Facebook. Questo perché Facebook espone molti più dettagli sul suo sistema pubblicitario agli inserzionisti che agli utenti. In effetti, consente agli inserzionisti di apprendere più informazioni che hanno dedotto o acquistato su un utente di quelle a cui l'utente stesso non consentirà l'accesso.³⁵⁹ L'esistenza delle categorie di targeting antisemite generate automaticamente, menzionate sopra, è stata scoperta utilizzando l'interfaccia dell'inserzionista. È stato riscontrato che la pubblicazione di annunci su Facebook introduce disparità demografiche dovute sia a effetti di mercato che a effetti di ottimizzazione dell'efficacia.²⁷⁴ Per ribadire, ciò significa che anche se l'inserzionista non targetizza esplicitamente un annuncio in base (ad esempio) al sesso, potrebbe esserci un criterio sistematico di genere, distorcere il pubblico dell'annuncio. Gli effetti di ottimizzazione sono abilitati dall'analisi di Facebook dei contenuti degli annunci. È interessante notare che questo include l'analisi delle immagini, che i ricercatori hanno rivelato utilizzando la tecnica intelligente di pubblicare annunci con contenuti trasparenti che sono invisibili agli esseri umani ma che hanno comunque un effetto sulla pubblicazione degli annunci.²⁷⁴

Considerazioni sull'equità nella progettazione dei mercati online

Le piattaforme online per il ride hailing, gli alloggi a breve termine e il lavoro freelance (concorso) sono diventate importanti negli anni 2010: esempi degni di nota sono Uber, Lyft, Airbnb e TaskRabbit. Sono obiettivi importanti per lo studio dell'equità perché hanno un impatto diretto sui mezzi di sussistenza e sulle opportunità delle persone. Metteremo da parte alcuni tipi di mercati dalla nostra discussione. Le app di appuntamenti online condividono alcune somiglianze con questi mercati, ma richiedono un'analisi completamente separata perché le norme che governano il romanticismo sono diverse da quelle che governano il commercio e l'occupazione.³⁶⁰ Poi ci sono mercati per beni come Amazon ed eBay. In questi mercati le caratteristiche dei partecipanti sono meno salienti degli attributi del prodotto, quindi la discriminazione è meno preoccupante (il che non vuol dire che sia inesistente³⁶¹).

A differenza dei settori studiati finora, il machine learning non è una componente fondamentale degli algoritmi nei mercati online. (Tuttavia, lo consideriamo di portata a causa del nostro ampio interesse per il processo decisionale e l'equità, piuttosto che solo per l'apprendimento automatico.) Pertanto le preoccupazioni sull'equità riguardano meno l'addestramento dei dati o degli algoritmi; il problema molto più serio è la discriminazione da parte di acquirenti e venditori. Ad esempio, uno studio ha scoperto che gli autisti di Uber disattivavano l'app nelle aree in cui non volevano far salire i passeggeri.³⁶²

I metodi per rilevare la discriminazione nei mercati online sono abbastanza simili a quelli utilizzati in contesti tradizionali quali l'alloggio e l'occupazione; è stata utilizzata una combinazione di studi di audit e metodi osservativi. Un esempio degnio di nota è un esperimento sul campo condotto da Edelman, Luca e Svirsky contro Airbnb.³⁶³ Gli autori hanno creato account di ospiti falsi i cui nomi indicavano razza (afro-americano o bianco) e sesso (femmina o maschio), ma per il resto erano identici. Sono stati utilizzati venti nomi diversi: cinque per ciascuna combinazione di razza e sesso. Attraverso questi account hanno quindi contattato gli host di 6.400 annunci in cinque città per chiedere informazioni sulla disponibilità. Hanno riscontrato una probabilità del 50% di accettazione delle richieste degli ospiti

con nomi dal suono bianco, rispetto al 42% per gli ospiti con nomi dal suono afroamericano . L'effetto era persistente indipendentemente dalla razza, dal sesso e dall'esperienza dell'host sulla piattaforma, nonché dal tipo di annuncio (prezzo alto o basso; proprietà intera o condivisa) e dalla diversità del quartiere. Tieni presente che gli account non avevano immagini del profilo; Se la deduzione della razza da parte degli host avviene in parte in base all'apparenza, un progetto di studio che vari le immagini del profilo degli account potrebbe avere un effetto maggiore.

Rispetto ai contesti tradizionali, alcuni tipi di dati osservativi sono facilmente disponibili su piattaforme online, che possono essere utili al ricercatore. Nello studio di cui sopra, la disponibilità pubblica delle recensioni delle proprietà elencate si è rivelata utile. Non è stato essenziale per la progettazione dello studio, ma ha consentito un interessante controllo di validità. Quando l'analisi è stata limitata al 29% degli host del campione che avevano ricevuto almeno una recensione da un ospite afroamericano, la disparità razziale nelle risposte è diminuita drasticamente. Se i risultati dello studio fossero il risultato di una stranezza del disegno sperimentale, piuttosto che di un'effettiva discriminazione razziale da parte degli host di Airbnb, sarebbe difficile spiegare perché l'effetto scomparirebbe per questo sottoinsieme di host. Ciò supporta la validità esterna dello studio.

Oltre alla discriminazione da parte dei partecipanti, un altro problema di equità con cui molti mercati online devono confrontarsi sono le differenze geografiche nell'efficacia. Uno studio di TaskRabbit e Uber ha rilevato che i quartieri ad alta densità di popolazione e ad alto reddito ricevono i maggiori benefici dalla sharing economy.³⁶⁴ A causa della pervasiva correlazione tra povertà e razza/etnicità, questi si traducono anche in disparità razziali. Nell'area di Chicago, dove è stato condotto questo studio, i quartieri neri e latini hanno una densità di popolazione inferiore, aggravando ulteriormente questo effetto.

Naturalmente, le disparità geografiche e strutturali in questi mercati non sono causate dalle piattaforme online, e senza dubbio esistono analogie offline come il passaparola . In effetti, l'entità della discriminazione razziale è molto maggiore in scenari come quello di fermare un taxi per strada³⁶⁵ rispetto alle interazioni mediate dalla tecnologia. Tuttavia, rispetto ai mercati regolati dalla legge antidiscriminatoria, come gli hotel, la discriminazione nei mercati online è più grave. In ogni caso, la natura formalizzata delle piattaforme online facilita gli audit. Inoltre, la natura centralizzata di queste piattaforme rappresenta una potente opportunità per interventi sull'equità.

Esistono molti modi in cui le piattaforme possono utilizzare la progettazione per ridurre al minimo la capacità degli utenti di discriminare (ad esempio trattenendo informazioni sulle controparti) e l'impulso a discriminare (ad esempio rendendo le caratteristiche dei partecipanti meno salienti rispetto alle caratteristiche del prodotto nell'interfaccia).³⁶⁶ Non è possibile per le piattaforme assumere una posizione neutrale nei confronti della discriminazione da parte dei partecipanti: anche le scelte fatte senza esplicito riguardo alla discriminazione possono influenzare la misura in cui gli atteggiamenti pregiudizievoli degli utenti si traducono in comportamenti discriminatori.

Come esempio concreto di decisioni progettuali volte a mitigare la discriminazione, gli autori dello studio Airbnb raccomandano che la piattaforma nasconde le informazioni sugli ospiti agli host prima della prenotazione. (Si noti che i servizi di ride hailing trattengono le informazioni sui clienti. I servizi di car pooling, d'altro canto, consentono agli utenti di visualizzare i nomi quando selezionano le corrispondenze; non sorprende che ciò consenta la discriminazione contro

nic minoranze.³⁶⁷⁾ Gli autori dello studio sulle disuguaglianze geografiche suggeriscono, tra gli altri interventi, che i servizi di ride hailing forniscano un punteggio di "reputazione geografica" agli autisti per combattere il fatto che gli autisti spesso percepiscono erroneamente i quartieri come più pericolosi di quanto non siano. Sono.

Meccanismi di discriminazione

Abbiamo esaminato una serie di studi sull'individuazione delle ingiustizie nei sistemi algoritmici. Facciamo il punto.

Nel capitolo introduttivo abbiamo discusso, ad alto livello, i diversi modi in cui potrebbero verificarsi ingiustizie nei sistemi di apprendimento automatico. Qui vediamo che le fonti e i meccanismi specifici dell'ingiustizia possono essere complessi e specifici del dominio.

I ricercatori hanno bisogno di comprendere il settore per formulare e testare in modo efficace ipotesi sulle fonti e sui meccanismi di ingiustizia.

Ad esempio, si consideri lo studio dei sistemi di classificazione di genere discusso sopra. È facile intuire che set di dati di addestramento non rappresentativi abbiano contribuito alle disparità di accuratezza osservate, ma non rappresentativi in che modo? Un articolo di follow-up di Muthukumar et al. ha considerato questa questione.³⁶⁸ Ha analizzato diversi classificatori di genere all'avanguardia (in un ambiente di laboratorio, in contrapposizione ai test sul campo delle API commerciali nel documento originale) e ha sostenuto che la sottrarappresentazione delle tonalità della pelle più scure nei dati di addestramento non è una ragione per la disparità osservata.

Invece, un meccanismo suggerito dagli autori si basa sul fatto che molti set di dati di addestramento di volti umani comprendono foto di celebrità.⁶ Hanno scoperto che le foto delle celebrità femminili hanno un trucco più evidente rispetto alle foto delle donne in generale. Ciò ha portato i classificatori a utilizzare il trucco come indicatore del genere in un modo che non era generalizzato al resto della popolazione.

Ipotesi leggermente diverse possono produrre conclusioni molto diverse, soprattutto in presenza di interazioni complesse tra produttori di contenuti, consumatori e piattaforme. Ad esempio, uno studio di Robertson et al. hanno testato affermazioni di parzialità da parte dei motori di ricerca, così come affermazioni correlate secondo cui i motori di ricerca restituiscono risultati che rafforzano le opinioni esistenti degli utenti (l'ipotesi della "bolla di filtro").³⁷⁰ I ricercatori hanno reclutato partecipanti con opinioni politiche diverse, raccolto risultati di ricerca di Google su un argomento politico sia nelle finestre standard che in quelle di navigazione in incognito dai computer di quei partecipanti, e ha scoperto che i risultati di ricerca standard (personalizzati) non erano più partigiani di quelli in incognito (non personalizzati), apparentemente trovando prove contro l'affermazione che la ricerca online rafforza le convinzioni esistenti degli utenti .

Questa scoperta è coerente con il fatto che Google non personalizza i risultati di ricerca se non in base alla posizione dell'utente e alla cronologia immediata (10 minuti) delle ricerche. Ciò è noto per ammissione dello stesso Google³⁷¹ e da ricerche precedenti.³⁷² Tuttavia, un'ipotesi più plausibile per l'effetto bolla di filtro nella ricerca arriva

⁶Questa sovrarappresentazione è dovuta al fatto che le foto delle celebrità sono più facili da raccogliere pubblicamente e si ritiene che le celebrità abbiano indebolito i diritti alla privacy a causa dell'interesse pubblico concorrente nelle loro attività. Tuttavia, per un contrappunto, vedi.³⁶⁹

da uno studio qualitativo di Francesca Tripodi.³⁷³ Semplificato un po' per i nostri scopi, funziona come segue: quando si svolge un evento con significato politico, i principali influencer (politici, organi di informazione partigiani, gruppi di interesse, forum politici) costruiscono rapidamente le proprie narrazioni di l'evento. Tali narrazioni raggiungono selettivamente i rispettivi pubblici partigiani attraverso reti di informazione partigiane.

Queste persone poi si rivolgono ai motori di ricerca per saperne di più o per "verificare i fatti". Fondamentalmente, tuttavia, utilizzano termini di ricerca diversi per riferirsi allo stesso evento, riflettendo le diverse narrazioni a cui sono stati esposti.⁷ I risultati per questi diversi termini di ricerca sono spesso nettamente diversi, perché i produttori di notizie e commenti selettivamente e strategicamente soddisfare i partigiani utilizzando queste stesse narrazioni. Pertanto, le convinzioni dei ricercatori vengono rafforzate. Si noti che questo meccanismo di produzione di bolle di filtro funziona in modo efficace anche se l'algoritmo di ricerca stesso è discutibilmente neutrale.⁸ Un ultimo

esempio per rafforzare il fatto che i meccanismi di produzione di disparità possono essere sottili e che è necessaria esperienza nel settore per formulare la giusta ipotesi: un'un'indagine condotta dai giornalisti ha rilevato che staples.com mostrava prezzi scontati ai privati in alcuni codici postali; questi codici postali erano, in media, più ricchi.³⁷⁵ Tuttavia, la regola effettiva dei prezzi che spiegava la maggior parte della variazione, come riportato, era che se c'era un negozio fisico di un concorrente situato entro 20 miglia circa dalla posizione dedotta del cliente, allora il cliente vedrebbe uno sconto!

Presumibilmente questa strategia ha lo scopo di dedurre il prezzo di riserva o la disponibilità a pagare del cliente. Per inciso, questo è un tipo simile di "discriminazione statistica" come osservato nello studio sulla discriminazione nelle vendite di automobili discusso all'inizio di questo capitolo.

Criteri di equità negli audit algoritmici

Mentre i meccanismi di ingiustizia sono diversi nei sistemi algoritmici, i criteri di equità applicabili sono gli stessi per il processo decisionale algoritmico come per altri tipi di processo decisionale. Detto questo, alcuni concetti di equità sono più spesso rilevanti, e altri meno, nel processo decisionale algoritmico rispetto al processo decisionale umano. Sul punto proponiamo alcune osservazioni selezionate.

L'equità come cecità è vista meno spesso negli studi di audit dei sistemi algoritmici; tali sistemi sono generalmente progettati per essere ciechi rispetto agli attributi sensibili. Inoltre, le preoccupazioni sull'equità spesso nascono proprio dal fatto che la cecità generalmente non è un intervento efficace sull'equità nell'apprendimento automatico. Due eccezioni sono il targeting degli annunci e i mercati online (dove le decisioni non cieche vengono infatti prese dagli utenti e non dalla piattaforma).

⁷Ad esempio, nel 2017, il presidente degli Stati Uniti Donald Trump ha chiesto alla National Football League di licenziare i giocatori che si erano impegnati in una protesta politica molto pubblicizzata durante le partite. Le narrazioni opposte di questo evento erano che il numero di spettatori della NFL era diminuito a causa delle proteste dei fan contro le azioni dei giocatori, o che era aumentato nonostante le proteste. I termini di ricerca che riflettono queste opinioni potrebbero essere "rating NFL in ribasso" rispetto a "rating NFL in aumento".

⁸Ma vedere 374 ("Data Void Type #4: Fragmented Concepts") per un argomento secondo cui la decisione dei motori di ricerca di non comprimere i concetti correlati contribuisce a questa frammentazione.

Ingiustizia come arbitrarietà. Esistono grossomodo due sensi in cui il processo decisionale potrebbe essere considerato arbitrario e quindi ingiusto. Il primo è quando le decisioni vengono prese in base a un capriccio piuttosto che a una procedura uniforme. Poiché il processo decisionale automatizzato comporta uniformità procedurale, questo tipo di preoccupazione generalmente non è rilevante.

Il secondo senso di arbitrarietà si applica anche quando esiste una procedura uniforme , se tale procedura si basa sulla considerazione di fattori ritenuti irrilevanti , sia statisticamente che moralmente. Poiché l'apprendimento automatico eccelle nel trovare correlazioni, identifica comunemente fattori che sembrano sconcertanti o palesemente inaccettabili . Ad esempio, nei test attitudinali come il Graduate Record Examination, i saggi vengono valutati automaticamente. Sebbene l'e-rater e gli altri strumenti utilizzati a questo scopo siano soggetti a controlli di validazione e risultino funzionare in modo simile ai valutatori umani su campioni di saggi reali, possono essere ingannati nel dare punteggi perfetti a parole senza senso generate dalla macchina. Ricordiamo che non esiste un criterio semplice che ci permetta di valutare se una caratteristica è moralmente valida (Capitolo 2), e questa questione deve essere discussa caso per caso.

Problemi più seri sorgono quando i classificatori non sono nemmeno sottoposti ad adeguati controlli di validità. Ad esempio, esistono numerose aziende che affermano di prevedere l'idoneità dei candidati per un lavoro sulla base di test della personalità, del linguaggio del corpo e di altre caratteristiche presenti nei video.³⁵⁰ Non esiste alcuna prova sottoposta a revisione paritaria che la performance lavorativa sia prevedibile utilizzando questi fattori, e nemmeno base per tale convinzione. Pertanto, anche se questi sistemi non producono disparità demografiche, sono ingiusti nel senso che sono arbitrari: i candidati che ricevono una decisione sfavorevole non dispongono del giusto processo per comprendere le basi della decisione, contestarla o determinare come aumentare le loro possibilità di ottenere successo.

I criteri di equità osservativa, tra cui la parità demografica, la parità del tasso di errore e la calibrazione, hanno ricevuto molta attenzione negli studi sull'equità algoritmica. La comodità ha probabilmente giocato un ruolo importante in questa scelta: questi parametri sono facili da raccogliere e semplici da riportare senza necessariamente collegarli a nozioni morali di equità. Ribadiamo la nostra cautela riguardo all'uso eccessivo di nozioni basate sulla parità; raramente la parità dovrebbe diventare un obiettivo a sé stante. Come minimo, è importante comprendere le fonti e i meccanismi che producono le disparità, nonché i danni che ne derivano prima di decidere gli interventi appropriati.

Danni rappresentazionali. Tradizionalmente, i dati allocativi e rappresentazionali venivano studiati in letterature separate, riflettendo il fatto che sono per lo più visti in sfere separate della vita (ad esempio, la discriminazione abitativa rispetto agli stereotipi nella pubblicità). Molti sistemi algoritmici, invece, sono in grado di generare entrambi i tipi di dati. Il fallimento del riconoscimento facciale per le persone dalla pelle scura è umiliante, ma potrebbe anche impedire a qualcuno di accedere a un dispositivo digitale o entrare in un edificio che utilizza la sicurezza biometrica.

Flusso informativo, correttezza, privacy

Una nozione chiamata flusso di informazioni è presente frequentemente negli audit algoritmici. Tale criterio prevede che le informazioni sensibili relative ai soggetti non circolino da un sistema informativo all'altro, o da una parte di un sistema all'altra. Ad esempio, un sito web sanitario può promettere che l'attività dell'utente, come ricerche e clic, non viene condivisa con terzi come le compagnie di assicurazione (poiché ciò potrebbe portare a effetti potenzialmente discriminatori sui premi assicurativi). Può essere visto come una generalizzazione della cecità: mentre la cecità significa non agire sulle informazioni sensibili disponibili, limitare il flusso di informazioni garantisce che le informazioni sensibili non siano disponibili per agire in primo luogo.

Esiste un potente test per verificare le violazioni dei vincoli del flusso di informazioni, che chiameremo test contraddittorio.³⁵⁸ Non rileva direttamente il flusso di informazioni, ma piuttosto le decisioni prese sulla base di tali informazioni. È potente perché non richiede la specifica di una variabile target, il che riduce al minimo la conoscenza del dominio richiesta al ricercatore. Per illustrare, rivisitiamo l'esempio del sito web sanitario. La prova contraddittoria funziona come segue:

1. Creare due gruppi di utenti simulati (A e B), ovvero bot, che siano identici tranne per il fatto che gli utenti del gruppo A, ma non del gruppo B, navigano nel sito Web sensibile in questione.
2. Chiedere a entrambi i gruppi di utenti di navigare su altri siti Web che si ritiene pubblichino annunci di compagnie assicurative, o di personalizzare i contenuti in base agli interessi degli utenti, o in qualche modo adattare i contenuti agli utenti in base alle informazioni sanitarie. Questo è il punto chiave: il ricercatore non ha bisogno di ipotizzare un meccanismo attraverso il quale si ottengono risultati potenzialmente ingiusti, ad esempio quali siti web (o terze parti) potrebbero ricevere dati sensibili, se la personalizzazione potrebbe assumere la forma di pubblicità, prezzi o altro. aspetto del contenuto.
3. Registrare il contenuto delle pagine web viste da tutti gli utenti nel passaggio precedente.
4. Addestrare un classificatore binario per distinguere tra le pagine web incontrate dagli utenti del gruppo A e quelle incontrate dagli utenti del gruppo B. Utilizzare la convalida incrociata per misurarne l'accuratezza.
5. Se il vincolo del flusso di informazioni è soddisfatto (ovvero, il sito Web sanitario non ha condiviso alcuna informazione dell'utente con terze parti), i siti Web visitati nella fase 2 sono ciechi rispetto alle attività dell'utente nella fase 1; quindi i due gruppi di utenti sembrano identici e non c'è modo di distinguere sistematicamente il contenuto visto dal gruppo A da quello visto dal gruppo B. L'accuratezza del test del classificatore non dovrebbe superare significativamente 2 . Il test¹ di permutazione può essere utilizzato per quantificare la probabilità che l'accuratezza osservata del classificatore (o migliore) possa essersi verificata per caso se in realtà non esiste alcuna differenza sistematica tra i due gruppi.³⁷⁶

Ci sono ulteriori sfumature relative alla corretta randomizzazione e ai controlli, per i quali rimandiamo il lettore allo studio di Datta, Tschantz e Datta.³⁵⁸ Si noti che se il test contraddittorio non riesce a rilevare un effetto, ciò non significa che il vincolo del flusso di informazioni è soddisfatto. Si noti inoltre che la prova contraddittoria non è in grado di

misurare la dimensione dell'effetto. Una tale misurazione sarebbe comunque priva di significato, poiché l'obiettivo è rilevare il flusso di informazioni e qualsiasi effetto sul comportamento osservabile del sistema ne è semplicemente un proxy.

Questa visione del flusso di informazioni come una generalizzazione della cecità rivela un'importante connessione tra privacy ed equità. Molti studi basati su questo principio possono essere visti come indagini sulla privacy o sull'equità. Ad esempio, uno studio ha scoperto che Facebook richiede numeri di telefono agli utenti con lo scopo dichiarato di migliorare la sicurezza dell'account, ma utilizza tali numeri per il targeting degli annunci.³⁷⁷ Questo è un esempio di flusso di informazioni riservate da una parte all'altra del sistema. Un altro studio ha utilizzato il retargeting pubblicitario – in cui le azioni intraprese su un sito web, come la ricerca di un prodotto, danno come risultato la pubblicità di quel prodotto su un altro sito web – per dedurre lo scambio di dati degli utenti tra società pubblicitarie.³⁷⁸ Nessuno dei due studi ha utilizzato il test del contraddittorio.

Confronto dei metodi di ricerca

Per verificare l'equità degli utenti sulle piattaforme online, esistono due approcci principali: creare profili falsi e reclutare utenti reali come tester. Ognuno ha i suoi pro e contro. Entrambi gli approcci hanno il vantaggio, rispetto agli studi di audit tradizionali, di consentire una scala potenzialmente maggiore grazie alla facilità di creare account falsi o di reclutare tester online (ad esempio attraverso il crowd-sourcing).

Il dimensionamento è particolarmente rilevante per testare le differenze geografiche, data la portata globale di molte piattaforme online. In genere è possibile simulare utenti geograficamente dispersi manipolando i dispositivi di test per segnalare posizioni false. Ad esempio, la suddetta indagine sulle differenze di prezzo regionali su staples.com includeva in realtà una misurazione per ciascuno dei 42.000 codici postali negli Stati Uniti.³⁷⁹ Hanno ottenuto ciò osservando che il sito web memorizzava la posizione dedotta dell'utente in un cookie e procedendo a modificare a livello di codice il valore memorizzato nel cookie in ogni valore possibile.

Detto questo, nell'approccio del falso profilo sorgono comunemente ostacoli pratici. In uno studio, il numero di unità di prova era praticamente limitato dal requisito per ciascun account di avere una carta di credito distinta associata ad esso.³⁸⁰ Un altro problema è il rilevamento dei bot. Ad esempio, lo studio di Airbnb era limitato a cinque città, anche se inizialmente i ricercatori avevano pianificato di testarne di più, perché gli algoritmi di rilevamento dei bot della piattaforma sono entrati in azione durante il corso dello studio per rilevare e arrestare il modello anomalo di attività. È facile immaginare un risultato ancora peggiore in cui gli account rilevati come bot vengono in qualche modo trattati in modo diverso dalla piattaforma (ad esempio, i messaggi provenienti da tali account hanno maggiori probabilità di essere nascosti ai destinatari previsti), compromettendo la validità dello studio.

Come illustra questo esempio, il rapporto tra i ricercatori di audit e le piattaforme sottoposte ad audit è spesso conflittuale. Gli sforzi delle piattaforme per ostacolare i ricercatori possono essere tecnici ma anche legali. Molte piattaforme, in particolare Facebook, vietano sia la creazione di account falsi che l'interazione automatizzata nei loro Termini di servizio. L'etica della violazione dei Termini di servizio negli studi di audit è una questione di

dibattito in corso, parallelamente ad alcune delle discussioni etiche durante il periodo formativo degli studi di audit tradizionali. Oltre alle questioni etiche, i ricercatori corrono rischi legali quando violano i Termini di servizio. Infatti, in base a leggi come il Computer Fraud and Abuse Act degli Stati Uniti, è possibile che debbano affrontare sanzioni penali anziché semplicemente civili.

Rispetto all'approccio del profilo falso, il reclutamento di utenti reali consente un minore controllo sui profili, ma è in grado di catturare meglio la variazione naturale degli attributi e del comportamento tra i gruppi demografici. Pertanto, nessuno dei due progetti è sempre preferibile e sono in sintonia con diversi concetti di equità. Quando i tester vengono reclutati tramite crowdsourcing, il risultato è generalmente un campione di convenienza (ovvero il campione è sbilanciato verso persone facili da contattare), risultando in un campione non probabilistico (non rappresentativo). In genere non è possibile formare un tale gruppo di tester per eseguire un protocollo sperimentale; invece, tali studi tipicamente gestiscono l'interazione tra i tester e la piattaforma tramite strumenti software (ad esempio estensioni del browser) creati dal ricercatore e installati dal tester. Per ulteriori informazioni sulle difficoltà della ricerca utilizzando campioni non probabilistici, vedere il libro *Bit by Bit* di Matthew Salganik.³⁸¹

A causa delle gravi limitazioni di entrambi gli approcci, gli studi di laboratorio sui sistemi algoritmici sono comunemente visti. Il motivo per cui gli studi di laboratorio hanno valore è che, poiché i sistemi automatizzati sono completamente specificati utilizzando il codice, il ricercatore può sperare di simularli in modo relativamente fedele. Naturalmente esistono delle limitazioni: il ricercatore in genere non ha accesso ai dati di addestramento, ai dati di interazione dell'utente o alle impostazioni di configurazione. Ma la simulazione è un modo prezioso per gli sviluppatori di sistemi algoritmici di testare i propri sistemi, e questo è un approccio comune nel settore. Le aziende spesso arrivano al punto di rendere pubblici i dati anonimi sulle interazioni degli utenti in modo che ricercatori esterni possano condurre studi di laboratorio per sviluppare. Il Premio Netflix è un esempio lampante di tale diffusione di dati.³⁸² Finora, questi sforzi sono stati quasi sempre volti a migliorare l'accuratezza piuttosto che l'equità dei sistemi algoritmici.

Gli studi di laboratorio sono particolarmente utili per affrontare questioni che non possono essere studiate con altri metodi empirici, in particolare la dinamica dei sistemi algoritmici, ovvero la loro evoluzione nel tempo. Un risultato importante di questo tipo di studio è la quantificazione dei cicli di feedback nella polizia predittiva.^{30, 31} Un'altra intuizione è la crescente omogeneità dei modelli di consumo degli utenti nel tempo nei sistemi di

raccomandazione.³⁸³ Gli studi osservazionali e i criteri di equità osservativa continuano ad essere importanti tanto. Tali studi sono tipicamente condotti da sviluppatori di algoritmi o decisori, spesso in collaborazione con ricercatori esterni.^{384, 385} È relativamente raro che i dati osservativi siano resi pubblici. Una rara eccezione, il set di dati COMPAS, riguardava una richiesta del Freedom of Information Act.

Infine, vale la pena ribadire che gli studi quantitativi sono limitati in ciò che possono concettualizzare e misurare.³⁸⁶ Gli studi qualitativi ed etnografici sui decisori forniscono quindi una prospettiva complementare di inestimabile valore. Per illustrare, discuteremo uno studio di Passi e Barcas che riporta sei mesi di lavoro etnografico sul campo in un team aziendale di data science.¹⁷⁰ Il team ha lavorato su un progetto in

il dominio del finanziamento automobilistico che mirava a “migliorare la qualità” dei lead (i lead sono potenziali acquirenti di auto bisognose di finanziamenti che potrebbero essere convertiti in acquirenti effettivi attraverso il marketing). Dato un obiettivo di alto livello così amorfico, formulare un problema di scienza dei dati concreto e trattabile è un compito necessario e non banale, un compito ulteriormente complicato dalle limitazioni dei dati disponibili. Il documento documenta la sostanziale libertà nella formulazione dei problemi e mette in luce il processo iterativo utilizzato, che ha portato all'utilizzo di una serie di proxy per la qualità dei lead. Gli autori mostrano che diversi proxy hanno diverse implicazioni sull'equità: un proxy massimizzerebbe le opportunità di prestito delle persone e un altro allevierebbe i pregiudizi esistenti dei dealer, entrambi obiettivi di equità potenzialmente preziosi. Tuttavia, i data scientist non erano consapevoli delle implicazioni normative delle loro decisioni e non le hanno deliberate esplicitamente.

Guardando avanti

In questo capitolo abbiamo trattato i test tradizionali per la discriminazione e gli studi sull'equità di vari sistemi algoritmici. Insieme, questi metodi costituiscono un potente strumento per interrogare un singolo sistema decisionale in un singolo momento.

Ma ci sono altri tipi di domande sull'equità che possiamo porci: qual è l' effetto cumulativo della discriminazione subita da una persona nel corso della sua vita? Quali aspetti strutturali della società determinano ingiustizie? Non possiamo rispondere a una domanda del genere esaminando i singoli sistemi. Il prossimo capitolo riguarda l'ampliamento della nostra visione della discriminazione e l'utilizzo di tale prospettiva più ampia per studiare una serie di possibili interventi sull'equità.

Note del capitolo

Per comprendere gli studi sull'audit delle scienze sociali in modo più approfondito, vedere l'articolo di Devah Pager.³⁰² S. Michael Gaddis fornisce un'introduzione e un sondaggio più recenti.³⁸⁷

L'auditing dei sistemi algoritmici è un campo giovane e in rapida evoluzione: un articolo del 2014 ha lanciato un invito all'azione verso questo tipo di ricerca.^[^sandvig] La maggior parte degli studi che citiamo sono successivi a quel pezzo. Per una panoramica più recente focalizzata sui professionisti, Costanza-Chock, Raji e Buolamwini compilano le migliori pratiche per gli auditor e forniscono raccomandazioni per i policy maker sulla base di interviste con oltre 150 auditor.²⁸⁶ Vecchione, Levy e Baracas traggono lezioni per gli audit degli algoritmi e la giustizia da la storia degli audit nelle scienze sociali.³⁸⁸ Brundage e colleghi inseriscono gli audit di terze parti nel contesto di molti altri modi per supportare affermazioni verificabili sugli impatti dei sistemi di intelligenza artificiale.³⁸⁹

Per una trattazione approfondita della storia e della politica delle piattaforme informative, vedere The Master Switch di Tim Wu,³⁹⁰ The Politics of 'Platforms' and Custodians of the Internet di Tarleton Gillespie,³⁹¹ 392 e The New Governors di Kate Klonick.³⁹³

8

Una visione più ampia della discriminazione

I sistemi di apprendimento automatico non funzionano nel vuoto; vengono adottati in società che già presentano molti tipi di discriminazione intrecciati con sistemi di oppressione come il razzismo. Ciò è alla base delle preoccupazioni sull'equità nell'apprendimento automatico. In questo capitolo daremo uno sguardo sistematico alla discriminazione nella società. Questo ci fornirà un quadro più completo dei potenziali impatti dannosi dell'apprendimento automatico. Vedremo che, sebbene un'ampia varietà di interventi sull'equità siano possibili – e necessari – solo una piccola parte di essi si traduce in soluzioni tecniche.

Caso di studio: il divario salariale di genere su Uber

Utilizzeremo un documento che analizza il divario salariale di genere su Uber³⁹⁴ come un modo per applicare alcune lezioni dei due capitoli precedenti durante l'impostazione di alcuni dei temi di questo capitolo. Lo studio è stato redatto in collaborazione da attuali ed ex dipendenti di Uber.

Gli autori partono dall'osservazione che le autiste donne guadagnano il 7% in meno su Uber per ora attiva rispetto agli autisti uomini. Concludono che questo divario può essere spiegato da tre fattori: differenze di genere nelle scelte dei conducenti su dove guidare, maggiore esperienza degli uomini sulla piattaforma e tendenza degli uomini a guidare più velocemente. Scoprono che la discriminazione dei clienti e la discriminazione algoritmica non contribuiscono al divario. Prenderemo per ora colato le affermazioni tecniche dell'articolo, ma utilizzeremo il quadro critico che abbiamo introdotto per interpretare i risultati in modo molto diverso dagli autori.

Innanzitutto, comprendiamo i risultati in modo più dettagliato.

L'articolo analizza i dati osservativi sui viaggi negli Stati Uniti, principalmente a Chicago. Sopra, abbiamo disegnato un grafico causale che mostra quello che consideriamo essere il nucleo del modello causale studiato nell'articolo (gli autori non disegnano un grafico del genere e non pongono le loro domande in un quadro causale; abbiamo scelto di farlo quindi per scopi pedagogici). Un grafico completo sarebbe molto più grande della Figura; ad esempio, abbiamo omesso una serie di controlli aggiuntivi, come la razza, presentati nelle appendici.

Utilizzeremo questo grafico per descrivere i risultati. Ad alto livello, il grafico descrive una distribuzione congiunta i cui campioni sono i viaggi. Ad esempio, viaggi diversi corrispondenti allo stesso conducente avranno la stessa Residenza (a meno che il conducente

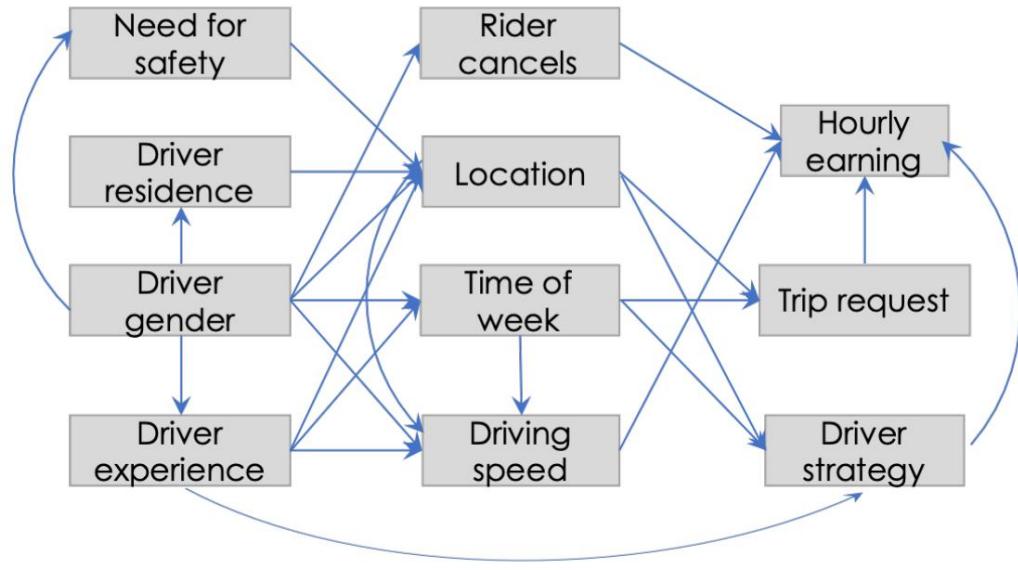


Figura 8.1: La nostra comprensione del modello causale隐式 nello studio Uber.

si sono spostati durante la loro permanenza sulla piattaforma), ma Esperienza diversa (misurata come numero di viaggi precedenti).

I guadagni orari degli autisti sono determinati principalmente dall'algoritmo che assegna le richieste di viaggio dai ciclisti agli autisti. L'assegnazione dipende dalla domanda, che a sua volta varia in base al luogo e all'ora della settimana (la variazione da una settimana all'altra è considerata rumore). L'algoritmo di Uber ignora gli attributi del conducente tra cui esperienza e sesso, quindi non ci sono frecce da questi nodi alla richiesta di viaggio. Inoltre, alcuni altri fattori potrebbero influenzare gli utili. I conducenti che guidano più velocemente completano più viaggi, i conducenti possono strategicamente accettare o annullare i viaggi e i ciclisti potrebbero discriminare annullando i viaggi dopo che il conducente ha accettato.

L'articolo utilizza una tecnica chiamata scomposizione di Gelbach per identificare l'effetto di ciascuna delle diverse variabili sulla retribuzione oraria. La decomposizione è un insieme di tecniche utilizzate in economia per quantificare il contributo di varie fonti a una differenza osservata nei risultati. Sebbene gli autori non eseguano un'inferenza causale, continueremo a parlare delle loro scoperte in termini causali per scopi pedagogici. La differenza non è rilevante per le questioni di alto livello che desideriamo sottolineare.

Gli autori ritengono che il divario di guadagno (cioè l'effetto del sesso del conducente sul guadagno orario) può essere interamente spiegato da percorsi che coinvolgono l'esperienza del conducente, la posizione e la velocità di guida. I percorsi con Cancellazione ciclista e Orario della settimana non hanno effetti significativi.

Gli autori esaminano ulteriormente l'effetto del genere sul luogo (ovvero la scelta di dove guidare) e scoprono che le donne hanno meno probabilità di guidare in aree meno sicure che risultano anche più redditizie. Poi scavano più a fondo e sostengono che questo effetto è operato quasi interamente dalle donne che risiedono in aree più sicure e scelgono di farlo

guidare in base a dove vivono.

I ritorni all'esperienza potrebbero operare in diversi modi. Gli autori non scompongono l'effetto ma suggeriscono diverse possibilità: la scelta di dove e quando guidare e altri elementi di strategia tra cui quali corse accettare. Una scoperta chiave dello studio è l'effetto del genere sull'esperienza. Gli uomini hanno meno probabilità di lasciare la piattaforma e guidare più ore durante ogni settimana in cui trascorrono sulla piattaforma, con conseguente grande differenza di esperienza. Non ci sono differenze di genere nell'imparare dall'esperienza: il comportamento dei conducenti uomini e donne cambia allo stesso ritmo per un dato numero di viaggi.

Il documento evidenzia questioni che possono essere studiate utilizzando dati osservativi ma non necessariamente con esperimenti sul campo (studi di audit). Uno studio sul divario retributivo di genere di Uber (sulla falsariga di quelli discussi nel capitolo precedente) potrebbe aver comportato la variazione del nome dell'autista per testare l'effetto sulla cancellazione e sulle valutazioni degli utenti. Un simile esperimento non sarebbe in grado di scoprire i numerosi altri percorsi attraverso i quali il genere influenza i guadagni. Uno studio di audit sarebbe più adatto per studiare la discriminazione dei conducenti nei confronti dei ciclisti, in parte perché i conducenti in questi sistemi esercitano una maggiore scelta nel processo di abbinamento rispetto ai ciclisti. In effetti, uno studio ha scoperto che gli autisti di UberX e Lyft discriminano gli utenti neri e le donne.³⁶⁵ I diagrammi causali in scenari realistici sono più complessi dei tipici

esempi da libri di testo. Ribadiamo che il grafico qui sopra è molto semplificato rispetto al grafico (implicito) nel documento. La stima nel documento procede come una serie di regressioni focalizzate in modo iterativo su piccole parti del grafico, piuttosto che su un'analisi dell'intero grafico in una sola volta. In qualsiasi esercizio complicato come questo, c'è sempre la possibilità di fattori confondenti non osservati.

Nonostante il numero di possibili effetti considerati nello studio, ne vengono esclusi molti altri. Ad esempio, alcuni conducenti potrebbero spostarsi per sfruttare il potenziale di guadagno. Ciò introduurrebbe un ciclo nel nostro grafo causale (Località → Residenza). Questo tipo di comportamento potrebbe sembrare improbabile per un singolo conducente, il che giustifica l'ignoranza di tali effetti nell'analisi. Nel corso del tempo, tuttavia, l'introduzione dei sistemi di trasporto ha il potenziale di rimodellare le comunità.^{395, 396} I metodi empirici odierni presentano limitazioni nella comprensione di questi tipi di fenomeni a lungo termine che implicano cicli di feedback.

Un'omissione più notevole nel documento riguarda l'effetto del genere del conducente sull'esperienza. Perché le donne abbandonano la piattaforma molto più frequentemente? Una delle ragioni potrebbe essere che devono affrontare maggiori molestie da parte dei ciclisti? Gli autori non sembrano prendere in considerazione questa questione.

Ciò porta alla nostra osservazione più saliente su questo studio: la definizione ristretta di discriminazione. In primo luogo, come notato, lo studio non considera che i differenziali tassi di abbandono potrebbero essere dovuti a discriminazione.¹ Ciò è particolarmente pertinente poiché il divario di genere nella retribuzione oraria è solo del 7% mentre il divario nel tasso di partecipazione è di 2,7!¹ Si potrebbe pensare che se ci fosse discriminazione tra i ciclisti, lo sarebbe

¹Ad esempio, gli autori affermano in astratto: I nostri risultati suggeriscono che, in un contesto di "gig economy" senza discriminazioni di genere e mercati del lavoro altamente flessibili, il costo opportunità relativamente elevato del tempo di lavoro non retribuito per le donne e le differenze basate sul genere nelle preferenze e vincoli possono sostenere un divario retributivo di genere.

essere più evidente nel suo effetto sui tassi di abbandono. Al contrario, l'unica via di discriminazione considerata nel documento riguarda un ciclista (presumibilmente misogino) che annulla una corsa, incorrendo in ritardi e potenzialmente sanzioni algoritmiche, basate esclusivamente sul sesso del conducente.

Inoltre, gli autori adottano una visione essenzialista della differenza di genere nella velocità media (ad esempio "gli uomini sono più tolleranti al rischio e aggressivi rispetto alle donne"). Potremmo chiederci quanto siano innate queste differenze, dato che nella società americana contemporanea, le donne possono essere soggette a sanzioni sociali quando vengono percepite come aggressive. Se questo è vero per le interazioni conducente-motociclista, allora le donne che guidano più velocemente degli uomini riceveranno valutazioni più basse con le conseguenti conseguenze negative. Questa è una forma di discriminazione da parte dei ciclisti.

Un'altra possibile visione della differenza di velocità, anch'essa non considerata dagli autori, è che gli autisti maschi in media forniscono una qualità di servizio inferiore a causa di un aumento del rischio di incidenti derivante dalla maggiore velocità (che crea anche esternalità negative per gli altri sulla strada). In quest'ottica, l'algoritmo di abbinamento di Uber discrimina le conducenti donne non tenendo conto di questa differenza.²

Infine, il documento non considera la discriminazione strutturale. Si scopre che le donne risiedono in quartieri meno redditizi e che il loro comportamento alla guida è modellato da considerazioni di sicurezza. Tuttavia, una comprensione più approfondita delle ragioni di queste differenze esula dallo scopo del documento. In effetti, le differenze di genere nei rischi per la sicurezza e nell'accessibilità economica dei quartieri residenziali possono essere viste come un esempio del maggiore onere che la società impone alle donne. In altre parole, Uber opera in una società in cui le donne subiscono discriminazioni e hanno un accesso ineguale alle opportunità, e la piattaforma perpetua tali differenze sotto forma di divario retributivo.³ Generalizziamo un po'. Esiste un'ampia serie di studi che cercano di spiegare le

ragioni delle disparità osservate nei salari o in altri risultati. Generalmente questi studi rilevano che l'effetto diretto del genere, della razza o di un altro attributo sensibile è molto minore dell'effetto indiretto. Spesso ciò porta a un acceso dibattito sulla questione se i risultati costituiscano o meno una prova di discriminazione o ingiustizia.

C'è spazio per opinioni diverse su questa questione. Gli autori dello studio Uber non hanno interpretato nessuno dei tre percorsi attraverso i quali il genere influisce sui guadagni – esperienza, velocità e posizione – come una discriminazione; abbiamo sostenuto che tutti e tre possono plausibilmente essere interpretati come discriminazione. Differenti quadri morali porteranno a risposte diverse. Le opinioni su queste questioni sono anche politicamente divise. Inoltre, studiosi di campi diversi spesso tendono a rispondere a queste domande in modo diverso (tra cui, notoriamente, le scienze sociali e l'economia³⁹⁷).

Certamente queste domande di definizione sono importanti. Tuttavia, forse il valore più grande degli studi sui meccanismi di discriminazione è che suggeriscono

² Se i conducenti assegnano valutazioni inferiori ai conducenti che guidano più velocemente a scapito della sicurezza, allora l'algoritmo di abbinamento tiene indirettamente conto delle considerazioni sulla sicurezza. Riteniamo improbabile che le valutazioni dei conducenti riflettano adeguatamente i rischi di eccesso di velocità, a causa di pregiudizi cognitivi. Dopotutto, è per questo che abbiamo bisogno di limiti di velocità invece di lasciarli decidere agli automobilisti.

³Vedere 364 per una discussione sui molti modi in cui si manifestano le disuguaglianze geografiche esistenti piattaforme di sharing economy tra cui Uber.

possibilità di intervento senza dover risolvere questioni di definizione. Osservando lo studio Uber da questo punto di vista, sono evidenti diversi interventi. Ricordiamo che esiste un'enorme disparità di genere nella velocità con cui i conducenti abbandonano la piattaforma. Uber potrebbe sollecitare e ascoltare più attivamente il feedback delle conducenti donne e utilizzare tale feedback per informare la progettazione dell'app. Ciò potrebbe portare ad interventi come rendere più semplice per i conducenti (e i ciclisti) denunciare le molestie e intraprendere azioni più forti in risposta a tali segnalazioni.

Per quanto riguarda la differenza di velocità, Uber potrebbe avvisare i conducenti che superano il limite di velocità o la cui velocità comporta un rischio di incidente previsto che supera una certa soglia (tale previsione è presumibilmente possibile dato l'accesso di Uber ai dati). Inoltre, Uber potrebbe utilizzare i suoi strumenti predittivi per istruire gli autisti sulla strategia, diminuendo il ritorno dell'esperienza per tutti gli autisti. Infine, i risultati danno anche maggiore urgenza agli sforzi strutturali per rendere i quartieri sicuri per le donne. Nessuno di questi interventi richiede un consenso sulla discriminazione o meno delle autiste di Uber .

Tre livelli di discriminazione

I sociologi organizzano la discriminazione in tre livelli: strutturale, organizzativo e interpersonale.^{325, 397} La discriminazione strutturale deriva dai modi in cui è organizzata la società, sia attraverso vincoli relativamente duri come leggi discriminatorie, sia attraverso vincoli più morbidi come norme e costumi. I fattori organizzativi operano a livello di organizzazioni o altre unità decisionali, come un'azienda che prende decisioni in materia di assunzioni. I fattori interpersonali si riferiscono agli atteggiamenti e alle convinzioni che determinano comportamenti discriminatori da parte degli individui.

Un modo separato per classificare la discriminazione è come diretta o indiretta. Per discriminazione diretta si intendono azioni o processi decisionali che fanno esplicito riferimento ad un attributo sensibile. Per discriminazione indiretta ci riferiamo ad azioni o processi decisionali che non fanno tale riferimento, ma svantaggiano uno o più gruppi. Il confine tra discriminazione diretta e indiretta è labile ed è meglio considerarlo come uno spettro piuttosto che come una categoria binaria.⁴

Tabella 8.1: Esempi di discriminazione organizzati in tre livelli e su uno spettro di direttività

Livello	Più diretto	Più indiretto
Strutturale	Leggi contro il matrimonio tra persone dello stesso sesso	Scuole meglio finanziate nelle aree più ricche e segregate
Mancanza organizzativa di alloggi per disabili		Assunzioni in rete
Interpersonale	Animus palese	Credenza nella necessità di una brillantezza innata (combinata con stereotipi di genere)

⁴Per i tentativi dei filosofi di formalizzare la distinzione, vedere.³⁹⁸ Per una trattazione tecnica degli effetti diretti e indiretti, fare riferimento al capitolo Causalità. Si veda anche;³⁹⁹ in particolare, il punto secondo cui “qualsiasi effetto diretto è in realtà un effetto indiretto se si approfondisce ulteriormente il meccanismo causale rilevante”.

Fattori strutturali

I fattori strutturali si riferiscono ai modi in cui è organizzata la società. Una legge che limita apertamente le opportunità per determinati gruppi è un esempio di fattore strutturale diretto. A causa delle varie rivoluzioni dei diritti in tutto il mondo, oggi ci sono meno di queste leggi rispetto al passato. Tuttavia, le leggi discriminatorie sono ben lungi dall'essere una cosa del passato. Ad esempio, nel 2021, solo 29 paesi riconoscono l'uguaglianza dei matrimoni.⁴⁰⁰ Inoltre, le leggi discriminatorie del passato hanno creato effetti strutturali che persistono ancora

oggi.⁴⁰¹ La discriminazione strutturale indiretta è pervasiva praticamente in ogni società. Ecco due esempi ben noti che riguardano gli Stati Uniti. Le leggi e le politiche sulla droga , nonostante siano apparentemente neutre, hanno l'effetto di influenzare in modo sproporzionato i gruppi minoritari, in particolare i neri.⁴⁰² Le scuole nei quartieri ad alto reddito tendono ad essere meglio finanziate (poiché le scuole pubbliche sono finanziate principalmente attraverso le tasse sulla proprietà) e attraggono più persone. insegnanti qualificati, trasmettendo un vantaggio educativo ai figli di genitori con redditi più alti.

Altri fattori sono ancora meno tangibili ma non meno gravi in termini di effetti, come le norme culturali e gli stereotipi. Nel caso di studio sui pregiudizi di genere nell'ammissione ai laureati di Berkeley nel Capitolo 5, abbiamo riscontrato l'ipotesi che gli stereotipi sociali influenzano le scelte di carriera delle persone in un modo che riproduce le disuguaglianze di genere nel reddito e nello status:

La distorsione dei dati aggregati deriva da . . . apparentemente da uno screening preliminare ai livelli precedenti del sistema educativo. Le donne vengono indirizzate dalla loro socializzazione e istruzione verso campi di studio universitari che sono generalmente più affollati, meno produttivi di titoli di studio completati e meno ben finanziati, e che spesso offrono prospettive di impiego professionale più povere.

Fattori organizzativi

I fattori organizzativi operano a livello di organizzazioni o unità decisionali : come sono strutturate, le regole e i processi decisionali che mettono in atto e il contesto in cui operano i singoli attori. Ancora una volta, questi si collocano in uno spettro tra diretto e indiretto.

La forma più diretta di discriminazione – escludere le persone dalla partecipazione esplicitamente basata sull'appartenenza ad un gruppo – è per lo più illegale nelle democrazie liberali. Tuttavia, pratiche come la mancanza di agevolazioni per i disabili e l'incapacità di combattere le molestie sessuali sono dilaganti. Una politica più indirettamente discriminatoria è l' utilizzo dei social network dei dipendenti durante le assunzioni, una pratica estremamente comune. Uno studio osservazionale ha rilevato che l'uso delle referenze dei dipendenti in aziende prevalentemente bianche riduceva la probabilità di un'assunzione di neri di quasi il 75% rispetto all'uso di annunci sui giornali.⁴⁰³ Lo studio ha controllato la segregazione spaziale, la segregazione occupazionale, la città e le dimensioni dell'azienda.

La discriminazione organizzativa può essere rivelata e affrontata a livello di una singola organizzazione, a differenza dei fattori strutturali (ad esempio, nessuna scuola singola è responsabile del fatto che gli insegnanti siano attratti da scuole situate in quartieri ad alto reddito).

Fattori interpersonali

I fattori interpersonali si riferiscono agli atteggiamenti e alle convinzioni che determinano comportamenti discriminatori da parte degli individui. A volte le persone possono discriminare a causa di un palese animus verso un certo gruppo, nel senso che il discriminatore non tenta di giustificarlo facendo alcun appello alla razionalità.

Più spesso, i meccanismi coinvolti sono relativamente indiretti. Uno studio del 2015 condotto da Leslie, Cimpian, Meyer e Freeland ha rilevato che i campi accademici in cui si ritiene che il successo sia guidato da una brillantezza innata mostrano una maggiore disparità di genere, cioè hanno meno donne.⁴⁰⁴ Gli autori propongono che la disparità sia causata dalla combinazione della convinzione nell'importanza della brillantezza innata insieme agli stereotipi sulla minore brillantezza innata nelle donne. Questa combinazione potrebbe quindi avere un impatto sulle donne nelle discipline che enfatizzano la brillantezza in due modi: o da parte dei praticanti di quelle discipline che mostrano pregiudizi contro le donne, o da parte delle donne che interiorizzano quegli stereotipi e si auto-selezionano da quelle discipline (o ottengono risultati più scarsi di quanto avrebbero altrimenti). volevo). Gli autori non progettano test per distinguere tra questi meccanismi concorrenti. Tuttavia, verificano se le disparità osservate potrebbero in alternativa essere causate da effettive differenze innate (piuttosto che da credenze in differenze innate) nelle capacità o attitudini, o nella volontà di lavorare per lunghe ore. Utilizzando vari proxy (come il punteggio GRE per l'abilità innata), sostengono che tali spiegazioni contrastanti non possono spiegare le differenze osservate.

Ci si potrebbe chiedere: non possiamo verificare le differenze innate in modo più rigoroso, ad esempio esaminando i bambini piccoli? Uno studio di follow-up ha mostrato che i bambini di sei anni tendono a interiorizzare stereotipi di genere sulla brillantezza innata, e questi stereotipi influenzano la loro scelta delle attività.⁴⁰⁵ Queste difficoltà suggeriscono la complessità di fondo del concetto di genere, che viene prodotto e rafforzato in parte attraverso questi stessi

stereotipi.⁴⁰⁶ Ricapitolando, abbiamo discusso della discriminazione strutturale, organizzativa e interpersonale , e del fatto che queste sono spesso indirette e pervasive. I tre livelli sono interconnessi: ad esempio, nel caso studio di Uber, le disuguaglianze strutturali non si perpetuano da sole, ma piuttosto attraverso decisioni organizzative; quelle decisioni in Uber sono prese da individui le cui visioni del mondo sono modellate dalla cultura. In altre parole, anche la discriminazione strutturale viene perpetrata attivamente e abbiamo collettivamente il potere di mitigarla e di invertire la rotta. Sarebbe un errore rassegnarsi a considerare la discriminazione strutturale semplicemente come il modo in cui va il mondo

Si noti che l'adozione di un processo decisionale statistico non è automaticamente una via d'uscita da nessuno di questi fattori, che operano per la maggior parte sullo sfondo e non in un singolo momento discreto del processo decisionale.

La persistenza e l'entità della disuguaglianza

L'uguaglianza formale ai sensi della legge affronta principalmente la discriminazione diretta e ha un effetto relativamente scarso sulla discriminazione indiretta, sia strutturale, organizzativa o interpersonale. Questa è una delle ragioni per cui la disuguaglianza può essere persistente nelle società che apparentemente promettono pari opportunità. Ecco due esempi lampanti di quanto a lungo le disuguaglianze possono mantenersi.

A partire dal 1609, furono istituite missioni gesuite nella regione Guaraní del Sud America, che si sovrappone ai moderni Argentina, Paraguay e Brasile. Oltre alla conversione religiosa, i missionari intrapresero sforzi educativi tra gli indigeni. Tuttavia, a causa degli sconvolgimenti politici in Spagna e Portogallo, le missioni terminarono bruscamente nel 1767-68 e i missionari furono espulsi. Quanto tempo dopo questa data dovremmo aspettarci che le disuguaglianze geografiche introdotte dalla presenza dei gesuiti persistano? Forse una o due generazioni? Sorprendentemente, si è scoperto che l'effetto dei Gesuiti sul livello di istruzione persiste 250 anni dopo: le aree più vicine a una precedente Missione hanno tassi di alfabetizzazione più alti del 10-15% e redditi più alti del 10%. Lo studio, di Felipe Valencia Caicedo, si avvale di un'idea intelligente per sostenere che i luoghi delle missioni erano essenzialmente casuali, rendendolo un esperimento naturale.⁴⁰⁷ Un altro studio sulla persistenza a lungo termine della disuguaglianza mostra gli effetti attuali di un sistema del lavoro forzato coloniale in Perù e Bolivia tra il 1573 e il 1812.⁴⁰⁸ Ulteriori prove della persistenza a lungo termine della disuguaglianza provengono dalla città di Firenze, sulla base di un insieme di dati unico contenente dati relativi alle tasse per tutti gli individui a partire dall'anno

^{1427.} Il documento di lavoro rileva che i cognomi associati a individui più ricchi nel set di dati sono associati a individui più ricchi oggi, seicento anni dopo.⁴⁰⁹ Sebbene questi siano solo alcuni esempi, la ricerca mostra che la persistenza della disuguaglianza nel corso delle generazioni lungo linee sociali e geografiche è la causa norma. Eppure non è molto apprezzato. Ad esempio, gli americani ritengono che un individuo nato nel quintile

più basso della distribuzione del reddito abbia una probabilità su 6 di raggiungere il quintile più alto, ma la probabilità osservata è 1 su 20.⁴¹⁰ La mobilità negli Stati Uniti è diminuita a partire dagli anni '80 ed è inferiore per i neri americani rispetto ai bianchi americani.⁴¹¹ Queste disuguaglianze sono significative sia per la loro entità che per la loro persistenza. Il reddito medio dei neri americani è circa il 65% di quello dei bianchi americani.⁴¹² La disuguaglianza di ricchezza è molto più grave: la ricchezza mediana delle famiglie nere è circa l'11% di quella delle famiglie bianche. Un'analisi dei dati combinata con simulazioni suggerisce che il

divario potrebbe non colmarsi mai senza interventi quali risarcimenti.⁴¹³ La maggior parte degli americani non è consapevole di questo divario: in media, gli intervistati stimano che la ricchezza di una tipica famiglia nera sia circa il 90% di quella di una famiglia nera, una tipica famiglia bianca.⁴¹⁴

Per quanto riguarda il genere, le donne che lavorano a tempo pieno, tutto l'anno, guadagnano l'80% di quanto guadagnano i loro colleghi maschi.⁴¹⁵ Esistono anche disuguaglianze geografiche. Ad esempio, i censimenti più ricchi e quelli più poveri degli Stati Uniti differiscono in termini di reddito medio di un fattore di circa 30.

Apprendimento automatico e discriminazione strutturale

Per un libro sull'apprendimento automatico, abbiamo trattato molti argomenti sulla discriminazione e la disuguaglianza nella società. C'è una ragione. Per comprendere l'equità non è sufficiente pensare al momento in cui si prende la decisione. Dobbiamo anche chiederci: quale impatto ha l'adozione dell'apprendimento automatico da parte dei decisori nei cicli duraturi di disuguaglianza strutturale nella società? Ci aiuta a fare progressi verso l'uguaglianza di opportunità, o altri ideali normativi, nel corso della vita delle persone? Ecco alcune osservazioni che possono aiutare a rispondere a queste domande.

I sistemi predittivi tendono a preservare vantaggi e svantaggi strutturali

I sistemi predittivi tendono a operare all'interno delle istituzioni esistenti. Quando tali istituzioni perpetuano la disuguaglianza dovuta a fattori strutturali, i sistemi predittivi non faranno altro che reificare quegli effetti, in assenza di un intervento esplicito. I sistemi predittivi tendono ad ereditare la discriminazione strutturale perché le funzioni obiettivo utilizzate nei modelli predittivi di solito riflettono gli incentivi delle organizzazioni che li implementano. Ad esempio, si consideri che uno studio del 2019 ha riscontrato forti pregiudizi razziali in un sistema utilizzato per identificare i pazienti ad alto rischio di esiti avversi per la salute, nel senso che ai pazienti neri sono stati assegnati punteggi inferiori rispetto ai pazienti bianchi ugualmente a rischio.³⁸⁴ Gli autori hanno scoperto che ciò è accaduto perché il modello è stato progettato per prevedere i costi sanitari anziché i bisogni, e il sistema sanitario spende meno per la cura dei pazienti neri rispetto ai pazienti bianchi anche quando hanno le stesse condizioni.

Supponiamo che un'azienda prenda decisioni di assunzione sulla base di un modello che prevede le prestazioni lavorative in base al livello di istruzione. Immaginate una società in cui gli studenti provenienti da famiglie a reddito più elevato, in media, hanno avuto migliori opportunità educative che si traducono in maggiori competenze lavorative. Questo non è un errore di misurazione dei dati che può essere corretto: il livello di istruzione predice realmente la prestazione lavorativa.

Pertanto, un sistema predittivo accurato classificherà in media i candidati con un reddito più elevato più in alto.

L'effetto strutturale di tali sistemi diventa chiaro quando immaginiamo che ogni datore di lavoro applichi considerazioni simili. I candidati con maggiori opportunità educative finiscono per ottenere posti di lavoro più desiderabili e redditi più alti. In altre parole, i sistemi predittivi hanno l'effetto di trasferire i vantaggi da una fase della vita a quella successiva e da una generazione a quella successiva.

Questo fenomeno si manifesta in modi meno evidenti. Ad esempio, il targeting degli annunci online si basa sul presupposto che le differenze nel comportamento passato tra gli utenti riflettono le differenze nelle preferenze. Ma potrebbero anche derivare da differenze nelle circostanze strutturali, e non c'è modo per i motori di prendere di mira di notare la differenza. Ciò aiuta a spiegare perché gli annunci, compresi gli annunci di lavoro, possono essere mirati in modi che rafforzano gli stereotipi e la discriminazione strutturale.⁴¹⁷ Questo

aspetto dei sistemi predittivi è amplificato dall'aggravamento dell'ingiustizia.^{418, 419} Cioè, gli individui sono soggetti a una serie di decisioni sulla corso della loro vita, e gli effetti di queste decisioni si accumulano e si aggravano nel tempo. Quando una persona riceve (o le viene negata) un'opportunità, è probabile che appaia di più (o

meno) qualificati al loro prossimo incontro con un sistema predittivo.

I sistemi di apprendimento automatico possono fare previsioni che si autoavverano

Supponiamo di scoprire che l'abilità negli scacchi è correlata alla produttività tra gli ingegneri del software. Ecco alcune possibili spiegazioni: 1. L'abilità negli scacchi rende un ingegnere del software migliore. 2. Ci sono abilità cognitive sottostanti che ci rendono migliori in entrambi. 3. I professori universitari mantengono stereotipi sulle abilità scacchistiche e sull'ingegneria del software e hanno indirizzato gli studenti bravi a scacchi verso corsi di informatica.

4. Le persone con più tempo libero sono state in grado di dedicarsi agli scacchi come hobby e di dedicare tempo al miglioramento delle proprie capacità di ingegneria del software.

L'apprendimento supervisionato standard non distingue tra questi percorsi causali. Indipendentemente dalla corretta spiegazione causale, una volta che un'ampia fascia di datori di lavoro inizia a utilizzare l'abilità scacchistica come criterio di assunzione, contribuisce alla perpetuazione della correlazione osservata. Questo perché i candidati più bravi negli scacchi avranno migliori opportunità per posizioni di ingegneria del software in questo mondo e queste opportunità consentiranno loro di sviluppare le proprie capacità di ingegneria del software.

L'apprendimento automatico automatizza la scoperta di correlazioni come quelle sopra. Quando utilizziamo queste correlazioni come criteri decisionali, alteriamo proprio i fenomeni che presumibilmente stiamo misurando. In altre parole, l'utilizzo di variabili non causali come criteri decisionali può conferire loro poteri causali nel tempo. Ciò non si limita all'apprendimento automatico: i sociologi riconoscono da tempo che gli stereotipi utilizzati per giustificare la discriminazione possono in realtà essere prodotti da tale discriminazione.⁴²⁰

I sistemi di raccomandazione algoritmica possono contribuire alla segregazione

Anche piccole preferenze per quartieri omogenei possono portare a drammatici effetti su larga scala. Nell'Appendice discutiamo un modello giocattolo di segregazione residenziale che mostra tali effetti. Ma che dire del mondo online, ad esempio dei social network online?

Il fenomeno per cui le persone fanno amicizia con altri simili (online o offline) è chiamato omofilia.

Agli albori dei social media, c'era la speranza – ora considerata ingenua – che nella sfera online non ci sarebbe stata segregazione grazie alla facilità con cui le persone potevano connettersi tra loro. Invece, osserviamo modelli simili di omofilia e segregazione online e offline. Ciò è dovuto in parte al fatto che le relazioni del mondo reale si riflettono online, ma in parte al fatto che la segregazione emerge attraverso le nostre preferenze e i nostri comportamenti online.⁴²¹

Con la maturazione dei social media, le preoccupazioni derivanti dall'omofilia si sono estese dalla segregazione demografica alle camere di risonanza ideologiche. I meccanismi causali dietro il discorso online polarizzato e il ruolo degli algoritmi di raccomandazione sono oggetto di ricerca e dibattito (vedi il capitolo Test), ma non c'è dubbio che i media online possano avere effetti strutturali.

L'apprendimento automatico può portare all'omogeneità del processo decisionale

Se un'azienda assume solo persone i cui nomi iniziano con determinate lettere dell'alfabeto, può sembrare assurdo ma non necessariamente motivo di allarme. Una delle ragioni alla base di questa intuizione è che ci aspettiamo che l'effetto di tali politiche idiosincratiche si annulli, dato che i candidati al lavoro hanno molte aziende a cui rivolgersi. Se, d'altro canto, ogni datore di lavoro adottasse una simile politica, l'esperienza di chi cerca lavoro diventerebbe radicalmente diversa.

L'apprendimento automatico si traduce in un processo decisionale più omogeneo rispetto ai capricci delle decisioni individuali. Gli studi sul comportamento umano mostrano che le decisioni umane contengono molto "rumore".⁵ La rimozione del rumore è una delle principali attrazioni del processo decisionale statistico. Ma ci sono anche dei rischi. Se il processo decisionale statistico portasse a prendere decisioni simili da parte di molti decisori, i pregiudizi altrimenti idiosincratici potrebbero essere amplificati e reificati al punto da creare impedimenti strutturali.⁵⁴ L'omogeneità può verificarsi in molti

modi. Ad alto livello, se molti sistemi di apprendimento automatico utilizzano gli stessi dati di addestramento e la stessa variabile target, faranno più o meno le stesse classificazioni, anche se gli algoritmi di apprendimento sono molto diversi. Intuitivamente, se così non fosse, si potrebbero fare classificazioni più accurate assemblando le loro previsioni. Per un chiaro esempio di previsioni omogenee nell'ambito della previsione degli esiti della vita, vedere *Fragile Families Challenge*.⁵⁰ In alternativa, molti decisori potrebbero utilizzare lo stesso sistema sottostante.

Kleinberg e Raghavan chiamano questa situazione monocultura algoritmica.⁴²³ Esistono aneddoti di persone in cerca di lavoro che sono state ripetutamente escluse dal lavoro sulla base di test della personalità, tutti offerti dallo stesso fornitore.⁴²⁴

Anche i singoli sistemi algoritmici possono avere un'influenza così sproporzionata nella società che le loro politiche possono avere effetti strutturali. L'esempio più ovvio sono i sistemi adottati dallo Stato, come un sistema di polizia predittiva che porta a un controllo eccessivo dei quartieri a basso reddito.

Ma sono le piattaforme private, soprattutto quelle su scala globale, dove questo effetto è stato più evidente. Prendiamo la moderazione dei contenuti: un piccolo numero di società di social media insieme determinano quali tipi di discorso possono far parte del discorso online tradizionale e quali comunità sono in grado di mobilitarsi online.

Le società che producono piattaforme sono state criticate per aver consentito contenuti che incitano alla violenza e, al contrario, per essere troppo zelanti nel deplatformare individui o gruppi.

In alcuni casi, le politiche della piattaforma sono modellate dalle capacità e dai limiti dell'apprendimento automatico.⁴²⁵ Ad esempio, gli algoritmi sono relativamente buoni nel rilevare la nudità ma relativamente scarsi nel rilevare il contesto. Aziende come Facebook hanno vietato ampiamente la nudità senza prestare molta attenzione al contesto, spesso eliminando opere d'arte e immagini storiche iconiche.

⁵Vedi.⁴²² L'articolo fa sia un'affermazione descrittiva sull'incoerenza delle decisioni umane sia un'affermazione normativa secondo cui un processo decisionale incoerente è un processo decisionale inadeguato. Quest'ultima affermazione può essere contestata in molti modi, uno dei quali verrà seguito qui.

L'apprendimento automatico sposta il potere

Come tutte le tecnologie, l'apprendimento automatico sposta il potere. Per renderlo più preciso, analizziamo l'adozione dell'apprendimento automatico da parte della burocrazia. Non intendiamo il termine burocrazia nel suo senso colloquiale e peggiorativo di un'agenzia governativa inefficiente e vincolata alle regole. Usiamo piuttosto il termine come fanno gli scienziati sociali: una burocrazia è un'entità pubblica o privata in cui lavoratori altamente qualificati chiamati burocrati, che operano in una struttura gerarchica, prendono decisioni in un modo che è vincolato da regole e politiche ma richiede anche il giudizio di esperti. . Le aziende, le università, gli ospedali, le forze di polizia e i programmi di assistenza pubblica sono tutte burocrazie a vari livelli. La maggior parte degli scenari decisionali che motivano questo libro sono situati nelle burocrazie.

Per comprendere l'effetto dell'adozione del machine learning, consideriamo cinque tipi di stakeholder: soggetti decisionali, persone che forniscono i dati di formazione, esperti di dominio, esperti di machine learning e policy maker. La nostra analisi si basa su un discorso di Pratyusha Kalluri.⁴²⁶

L'apprendimento automatico, come generalmente implementato oggi, sposta il potere dalle prime tre categorie. Rappresentando i soggetti decisionali come vettori di caratteristiche standardizzate, il processo decisionale statistico rimuove la loro agenzia e la capacità di difendere se stessi. In molti ambiti, in particolare nel sistema giudiziario, questa capacità è fondamentale per i diritti dei soggetti decisionali. Anche in un ambito relativamente meno importante come l'ammissione all'università, la dichiarazione personale fornisce questa capacità ed è una componente chiave della valutazione.

Le persone che forniscono dati di formazione possono avere conoscenza dell'attività da svolgere, ma forniscono solo il proprio comportamento come input al sistema (si pensi ai destinatari di posta elettronica che fanno clic sul pulsante "spam"). Il machine learning invece costruisce una forma di conoscenza in modo centralizzato. Al contrario, gli esperti di dominio apprendono in parte dalla conoscenza e dall'esperienza vissuta degli individui con cui interagiscono. Certo, esperti come i medici sono spesso criticati per aver svalutato la conoscenza e l'esperienza dei soggetti decisionali (i pazienti).⁴²⁷ Ma il fatto che un simile dibattito abbia luogo è la prova del fatto che i pazienti hanno almeno un certo potere nel sistema tradizionale. ⁴²⁸

Anche il ruolo degli esperti di settore è più limitato rispetto al tradizionale processo decisionale in cui prevalgono la discrezione e il giudizio di tali esperti.

Nell'apprendimento automatico supervisionato, le competenze del settore sono necessarie principalmente in due fasi: la formulazione del problema e del compito e l'etichettatura degli esempi di formazione. In pratica, le competenze nel settore spesso non sono apprezzate dagli sviluppatori di strumenti e quindi i ruoli degli esperti sono ancora più circoscritti. Ad esempio, uno studio ha rilevato che, sulla base di 68 interviste, "gli sviluppatori concepivano [gli esperti di dominio] come corrotti, pigri, non conformi e, a loro volta, perseguitavano la sorveglianza e la gamification per disciplinare i lavoratori e raccogliere dati di migliore qualità".⁴²⁹

Le implicazioni sull'equità di questo spostamento di potere sono complesse. Nelle burocrazie governative, il potere esercitato dai "burocrati di strada" come gli agenti di polizia e gli assistenti sociali dei servizi sociali – le persone che traducono la politica in decisioni individuali – può essere oggetto di abuso, e spesso si vede la rimozione della loro discrezione.

come intervento di equità. Tuttavia, la discrezione e l'intelligenza umana di questi decisori possono anche rappresentare un elemento vitale di promozione dell'equità a causa dell'esistenza di fattori attenuanti o circostanze nuove non riscontrate nei dati di formazione o contemplate nelle politiche esistenti.^{430, 59} E quando il sistema stesso è ingiusto, gli esseri umani incaricati di attuarlo possono costituire un'importante fonte di resistenza attraverso il mancato rispetto o la denuncia di irregolarità.

A differenza dei burocrati di strada, l'apprendimento automatico conferisce potere ai decisori politici o ai decisori centralizzati, ovvero quelli ai vertici della burocrazia.

Prendi in considerazione uno strumento di previsione del rischio utilizzato da un'agenzia per la protezione dei minori per filtrare le chiamate.

A seconda del budget dell'agenzia e di altri fattori, il decisore potrebbe voler filtrare una percentuale maggiore o minore di chiamate. Con uno strumento statistico, un simile cambiamento politico può essere implementato istantaneamente ed è enormemente più semplice dell'alternativa di riqualificare centinaia di operatori del caso per adattare le loro euristiche mentali. Questo è solo un esempio che illustra perché tali strumenti si sono rivelati così attraenti per coloro che decidono di implementarli.

Gli esperti di machine learning, ovviamente, tendono ad avere un ruolo centrale. I requisiti delle parti interessate devono essere tradotti in attuazione da questi esperti; intenzionalmente o meno, ci sono spesso divari sostanziali tra la politica desiderata e quella che viene realizzata nella pratica.²⁸⁰ In ogni sistema automatizzato, c'è qualcosa che si perde nella traduzione della politica dal linguaggio umano al codice informatico. Ad esempio, ci sono stati casi in cui il software ha calcolato erroneamente l'idoneità dei detenuti al rilascio anticipato, con conseguenze strazianti tra cui la detenzione in carcere troppo a lungo e il ritorno in carcere dopo essere stati rilasciati.^{431, 432} Ma nei classici sistemi automatizzati, queste lacune tendono a essere errori generalmente evidenti all'ispezione manuale (non che sia di conforto per coloro che vengono danneggiati). Ma quando è coinvolto il machine learning, il coinvolgimento dell'esperto è spesso necessario anche per riconoscere che qualcosa è andato storto. Questo perché la politica tende ad essere più ambigua (cosa significa "alto rischio"?) e perché le deviazioni dalla politica diventano evidenti solo in forma aggregata.

Inoltre, i decisori spesso abdicano il loro potere agli sviluppatori di strumenti, rendendoli ancora più potenti. Mulligan e Bamberger spiegano come le agenzie governative acquisiscono sistemi di apprendimento automatico attraverso processi di appalto, gli stessi processi utilizzati per garantire a un appaltatore di costruire un ponte.⁴³³ La mentalità degli appalti ignora il fatto che i prodotti risultanti vengono utilizzati per prendere decisioni consequenziali, ovvero elaborare politiche efficaci. . L'approvvigionamento enfatizza fattori come il prezzo e l'elusione del rischio piuttosto che la trasparenza o la supervisione del processo decisionale.

Interventi strutturali per il machine learning equo

Il fatto che l'apprendimento automatico possa contribuire alla discriminazione strutturale motiva la necessità di interventi di portata altrettanto ampia. Chiamiamo questi interventi strutturali: cambiare il modo in cui il machine learning viene costruito e distribuito. I cambiamenti che abbiamo in mente vanno oltre la portata di ogni singola organizzazione e richiedono un'azione collettiva. Questo potrebbe assumere la forma di un ampio

movimento sociale o altri collettivi tra cui comunità, lavoratori, ricercatori e utenti.

Riformare le istituzioni sottostanti

Un approccio è quello di concentrarsi sull'istituzione sottostante piuttosto che sulla tecnologia, e cambiarla in modo che sia meno incline ad adottare strumenti dannosi di apprendimento automatico. Ad esempio, spostare il focus del sistema di giustizia penale dall'inabilitazione alla riabilitazione potrebbe diminuire la domanda di strumenti di previsione del rischio.⁴³⁴ Molti studiosi e attivisti distinguono tra riforma e abolizione (a volte chiamata riforma non riformista), essendo l'abolizione una forma più radicale e transnazionale. - approccio formativo.^{435, 436, 193} Ai nostri fini, però, entrambi hanno l'effetto di centrare l'intervento sull'istituzione piuttosto che sulla tecnologia.

In molti ambiti, gli scopi e gli obiettivi stessi delle nostre istituzioni rimangono contestati. Ad esempio, quali sono gli obiettivi della polizia? Gli obiettivi comunemente accettati includono la deterrenza e la prevenzione della criminalità, la garanzia della sicurezza pubblica, la riduzione al minimo dei disordini e la consegna dei delinquenti alla giustizia; potrebbero anche includere sforzi più ampi per migliorare la salute e la vitalità delle comunità. L'importanza relativa di questi obiettivi varia tra le comunità e nel tempo. Pertanto, formulare le decisioni di allocazione della polizia come un problema di ottimizzazione, come fanno i sistemi di polizia predittiva, implica prendere posizione su queste questioni profondamente controverse.

La storia ci mostra che molte istituzioni che possono sembrare elementi fissi della società moderna, come l'istruzione superiore, hanno in realtà ridefinito ripetutamente i propri obiettivi e scopi per adattarsi a un mondo in cambiamento. In effetti, a volte lo scopo di tali cambiamenti è stato quello di discriminare in modo più efficace. All'inizio del XX secolo, le università americane d'élite si trasformarono dal considerare le dimensioni (in termini di iscrizioni) come una fonte di prestigio alla selettività. Una delle ragioni principali di questo cambiamento è stata quella di ridurre la percentuale crescente di studenti ebrei senza dover introdurre quote esplicite; la ritrovata missione di essere selettivi ha permesso loro di enfatizzare tratti come il carattere e la personalità nelle ammissioni, il che a sua volta ha consentito un ampio margine di discrezionalità. In effetti, questo sistema adottato da Harvard nel 1926 fu all'origine dell'approccio olistico alle ammissioni che continua a essere

controverso oggi, come spiega Jerome Karabel nel libro *The Chosen*.⁴³⁷ Alcuni studiosi sono andati oltre la posizione secondo cui l'intervento per affrontare i danni algoritmici dovrebbe concentrarsi sull'istituzione sottostante, e hanno sostenuto che l'adozione di un processo decisionale automatizzato consente effettivamente alle istituzioni resistenti di evitare le riforme necessarie. Virginia Eubanks esamina quattro programmi di assistenza pubblica per i poveri negli Stati Uniti: assistenza alimentare, Medicaid, senzatetto e bambini a rischio.⁴³⁸ In ciascun caso esistono criteri di ammissibilità gestiti automaticamente, alcuni dei quali utilizzano tecniche statistiche. Il libro documenta gli effetti dannosi di questi sistemi, compresi gli effetti punitivi su coloro ritenuti non idonei; l'impatto sproporzionato di tali oneri sulle persone di colore a basso reddito, in particolare sulle donne; la mancanza di trasparenza e l'apparente arbitrarietà delle decisioni; e il monitoraggio e la sorveglianza della vita dei poveri, necessari affinché questi sistemi funzionino.

Questi problemi possono essere risolvibili in una certa misura, ma Eubanks ha una critica più profonda: questi sistemi distraggono dall'obiettivo più fondamentale di sradicare la povertà ("Gestiamo i singoli poveri per sfuggire alla nostra responsabilità condivisa di sradicare la povertà"). In teoria i due approcci potrebbero coesistere. In pratica, sostiene Eubanks, questi sistemi legittimano l'idea che ci sia qualcosa che non va in alcune persone, nascondono il problema strutturale sottostante e favoriscono l'inazione.

Inoltre, comportano un costo monetario elevato che altrimenti potrebbe essere destinato a riforme più radicali.

Diritti della comunità

Le tecnologie dannose sono spesso giuridicamente giustificate nell'ambito di un quadro di notifica e consenso che si basa su una concezione individualistica dei diritti ed è mal attrezzato per affrontare i danni collettivi. Ad esempio, i dipartimenti di polizia ottengono filmati in massa dalle telecamere di sicurezza residenziali con il consenso dei residenti attraverso piattaforme centralizzate come Amazon Ring.⁴³⁸ Tuttavia, il consenso non è un controllo significativo in questo scenario, perché le persone che rischiano di essere danneggiate dagli abusi della polizia i filmati di sorveglianza – come manifestanti o membri di minoranze razziali che sono stati denunciati dalla polizia per "aver agito in modo sospetto" – non sono quelli per cui si chiede o si ottiene il consenso.

Questo divario è particolarmente rilevante nelle applicazioni di machine learning: anche se un classificatore è addestrato sui dati forniti con il consenso, può essere applicato a soggetti decisionali non consenzienti. Un'alternativa è quella di concedere a gruppi, come le comunità geografiche , il diritto di acconsentire o rifiutare collettivamente l'adozione di strumenti tecnologici. In risposta all'uso del riconoscimento facciale da parte della polizia, gli attivisti per le libertà civili hanno sostenuto il diritto della comunità di rifiutare tali strumenti; il successo di questa azione di sostegno ha portato a vari divieti e moratorie a livello locale.⁴³⁹ Al contrario, si consideri la pubblicità mirata online, un'altra tecnologia che ha dovuto affrontare un diffuso dissenso. In questo caso non esistono collettivi analoghi che possano organizzare una resistenza efficace, e quindi i tentativi di rifiutare la tecnologia hanno avuto molto meno successo.⁴⁴⁰

Al di là del consenso collettivo, un altro obiettivo dell'azione comunitaria è quello di ottenere un posto al tavolo nella progettazione di sistemi di machine learning come stakeholder e partecipanti le cui competenze ed esperienze vissute modellano la concezione e l'implementazione del sistema piuttosto che semplici fornitori di dati e soggetti decisionali.

Tra gli altri vantaggi, questo approccio renderebbe più semplice prevedere e mitigare i danni rappresentazionali, problemi come le categorie umilianti nei set di dati di visione artificiale o i risultati di ricerca di immagini che rappresentano stereotipi offensivi. Ma ci sono anche potenziali rischi per la progettazione partecipativa: potrebbe creare ulteriori oneri per i membri di comunità sottorappresentate e potrebbe fungere da cortina di fumo per le organizzazioni che si oppongono a cambiamenti significativi. È essenziale che la partecipazione sia riconosciuta come lavoro e sia equamente retribuita.⁴⁴¹

Regolamento

La regolamentazione che promuove l'apprendimento automatico equo può assumere la forma dell'applicazione delle leggi esistenti ai sistemi decisionali che incorporano l'apprendimento automatico, o leggi che affrontano specificamente l'uso della tecnologia e i danni che ne derivano. Esempi di questi ultimi includono i divieti di riconoscimento facciale sopra menzionati e le restrizioni al processo decisionale automatizzato ai sensi del Regolamento generale sulla protezione dei dati (GDPR) dell'Unione Europea . Entrambi i tipi di regolamentazione si stanno evolvendo in risposta alla rapida adozione dell'apprendimento automatico nei sistemi decisionali. La regolamentazione rappresenta un'importante opportunità di intervento strutturale per un apprendimento automatico equo. Tuttavia, a causa della tendenza della legge a concettualizzare la discriminazione in termini ristretti, il suo effetto pratico nel frenare il dannoso apprendimento automatico resta in gran parte da vedere.⁴⁴²

Il divario tra il ritmo di adozione dell'apprendimento automatico e il ritmo di evoluzione della legge ha portato a tentativi di autoregolamentazione: uno studio del 2019 ha rilevato 84 linee guida etiche sull'intelligenza artificiale in tutto il mondo.⁴⁴³ Tali documenti non hanno forza di legge ma tentano di definire norme per organizzazioni e/o singoli professionisti. Sebbene l'autoregolamentazione sia stata efficace in alcuni campi come la medicina, è dubbio che l'autoregolamentazione dell'IA possa affrontare i problemi spinosi che abbiamo identificato in questo capitolo.

In effetti, l'autoregolamentazione del settore mira generalmente a prevenire la regolamentazione vera e propria e i cambiamenti strutturali che potrebbe rendere necessari.⁶

Interventi della forza lavoro

Il machine learning trasferisce il potere agli esperti di machine learning, il che rende la forza lavoro del machine learning un importante luogo di intervento. Una serie di sforzi mira a consentire a più persone di beneficiare di preziose opportunità di lavoro nel settore⁴⁴⁵ e a combattere gli squilibri di potere all'interno della forza lavoro, in particolare tra esperti di tecnologia e coloro che svolgono altri ruoli come l'annotazione.⁴⁴⁶ Un'altra serie di gli sforzi mirano ad allineare gli usi del machine learning ai valori etici della forza lavoro del machine learning.

Il nascente movimento di sindacalizzazione nelle aziende tecnologiche sembra avere entrambi gli obiettivi.

Sebbene una forza lavoro più diversificata abbia un valore morale di per sé, è interessante chiedersi quale effetto abbia sull'equità dei prodotti risultanti. Uno studio sperimentale sui programmatore ha rilevato che il genere o la razza dei programmatore non influiva sulla produzione di codice distorto.⁴⁴⁷ Tuttavia, questo è uno studio di laboratorio e non dovrebbe essere visto come una guida agli effetti degli interventi strutturali. Ad esempio, un percorso causale attraverso il quale la diversità della forza lavoro potrebbe avere un impatto sui prodotti (non catturato nella progettazione dello studio) è che un team con una diversità di prospettive potrebbe essere più disposto a porre domande critiche sull'opportunità di costruire o implementare un prodotto .

Un altro intervento sulla forza lavoro riguarda l'istruzione e la formazione. L'educazione all'etica per gli studenti di informatica è in aumento e una raccolta del 2018 comprendeva oltre 200 corsi di questo tipo.⁴⁴⁸ Un dibattito di lunga data riguarda i meriti relativi dei corsi autonomi e l'integrazione dell'etica nell'informatica esistente.

⁶Per una critica più approfondita delle dichiarazioni di principi promosse dall'industria vedere.⁴⁴⁴

corsi.449 Organizzazioni professionali come l'Association for Computing Machinery (ACM) hanno codici etici da diversi decenni, ma non è chiaro se questi codici abbiano avuto un impatto significativo sui professionisti.

In molti campi professionali, compresi alcuni campi dell'ingegneria, le responsabilità etiche vengono rafforzate in parte attraverso la concessione di licenze ai professionisti. Professionisti come medici e avvocati devono padroneggiare un insieme di conoscenze professionali, compresi i codici etici, sono tenuti per legge a superare esami standardizzati prima di ottenere la licenza per esercitare e possono vedere revocata tale licenza se commettono trasgressioni etiche. Questo non è il caso dell'ingegneria del software. In ogni caso, gli standard di certificazione dell'ingegneria del software esistenti⁴⁵⁰ non hanno praticamente alcuna sovrapposizione con gli argomenti trattati in questo libro.

La comunità di ricerca

La comunità di ricerca sull'apprendimento automatico è un altro importante luogo di riforma e trasformazione. La spinta più significativa al cambiamento è stata la lotta continua per trattare argomenti di ricerca come l'equità, l'etica e la giustizia come legittimi e di prim'ordine.

Tradizionalmente, alcuni argomenti dell'apprendimento automatico, come gli algoritmi di ottimizzazione, sono stati considerati l'apprendimento automatico "core" o "reale", mentre altri argomenti, compresa la costruzione di set di dati, sono stati visti come periferici e meno intellettualmente seri. Birhane et al. hanno eseguito un'analisi testuale degli articoli presentati alle principali conferenze sull'apprendimento automatico, ICML e NeurIPS, e hanno scoperto che la maggior parte degli articoli si giustifica facendo appello a valori quali prestazioni e generalizzazione, e solo l'1% ha menzionato potenziali effetti negativi.⁴⁵¹

Alcuni altri dibattiti chiave: tutti i ricercatori nel campo dell'apprendimento automatico dovrebbero essere tenuti a riflettere sull'etica della loro ricerca?⁴⁵² C'è troppa attenzione alla correzione dei pregiudizi rispetto a questioni più profonde su potere e giustizia⁴⁵³? Come centrare le prospettive delle persone e delle comunità interessate dai sistemi di machine learning?

Qual è il ruolo della ricerca di settore sull'apprendimento automatico equo considerati i conflitti di interessi?

Interventi organizzativi per un processo decisionale più giusto

Gli interventi strutturali di cui abbiamo discusso sopra richiedono movimenti sociali o altre azioni collettive e si sono evoluti in un arco temporale che va da anni a decenni.

Questo non vuol dire che un'organizzazione debba alzare le mani e aspettare cambiamenti strutturali. Sono disponibili numerosi interventi per la maggior parte dei decisori . Questa sezione è una panoramica di quelli più importanti.

Mentre leggi, osserva che la maggior parte degli interventi tenta di migliorare i risultati per tutti i soggetti decisionali piuttosto che considerare l'equità come un compromesso inevitabile. Uno dei motivi per cui ciò è possibile è che molti di essi non sono operativi al momento della decisione. Si noti, inoltre, che la valutazione degli effetti degli interventi, sia rispetto all'equità che ad altri parametri, richiede generalmente un'inferenza causale. Infine , solo un piccolo sottoinsieme di potenziali interventi sull'equità può essere implementato

Type	Intervention	Example
Modifying the decision process	Reallocation	Group-specific decision thresholds
	Combatting interpersonal discrimination	Implicit bias training
	Formalization	Adopting statistical decision making
Before the decision	Procedural protections	Explanation and recourse
	Outreach	Sending mailers about scholarships
After the decision	Intervening on causal factors	Job training, preventive health
	Modifying the environment	Helping defendants show up to court

Figura 8.2: Una sintesi delle principali tipologie di interventi organizzativi

il quadro dell'apprendimento automatico. Gli altri si concentrano sulle pratiche organizzative o umane piuttosto che sul sottosistema tecnico coinvolto nel processo decisionale.

Ridistribuzione o riallocazione

Ridistribuzione e riallocazione sono termini che si riferiscono a interventi che modificano un processo decisionale per introdurre una preferenza esplicita per uno o più gruppi, solitamente gruppi considerati svantaggiati. Quando parliamo di interventi sull'equità, questo potrebbe essere il tipo che mi viene in mente più facilmente.

Quando applicati a problemi di selezione in cui esiste un numero relativamente statico di posti, come è tipico nelle assunzioni o nelle ammissioni al college, una plethora di interventi di equità algoritmica si riducono a diverse forme di riallocazione. Ciò include tecniche come l'aggiunta di un vincolo di equità alla fase di ottimizzazione o un aggiustamento post-elaborazione per migliorare i punteggi dei membri dei gruppi svantaggiati.

Ciò è vero indipendentemente dal fatto che l'obiettivo sia la parità demografica o qualsiasi altro criterio statistico.

La riallocazione è interessante perché non richiede una comprensione causale del motivo per cui è sorta la disparità. Per lo stesso motivo, la riallocazione è un intervento grossolano. È progettato per avvantaggiare un gruppo – e ha il vantaggio di fornire una misura di trasparenza consentendo una quantificazione del beneficio del gruppo – ma la maggior parte delle procedure di riallocazione non incorporano una nozione di merito dei membri all'interno di quel gruppo. Spesso la riallocazione si realizza attraverso una preferenza uniforme per i membri del gruppo svantaggiato. In alternativa, ciò può essere ottenuto armeggiando con l'obiettivo di ottimizzazione per incorporare una preferenza di gruppo. In questo approccio, la distribuzione dei frutti della riallocazione all'interno del gruppo è delegata al modello, che potrebbe finire per apprendere un'allocazione non intuitiva e non intenzionale (ad esempio, un sottogruppo intersezionale potrebbe finire per

svantaggiati rispetto ad una condizione di non intervento). Nella migliore delle ipotesi, i metodi di riallocazione mireranno a garantire che la classifica relativa all'interno dei gruppi rimanga invariata.

Per quanto cruda possa essere la riallocazione, un altro intervento con un compromesso ancora peggiore è quello di omettere dalla considerazione le caratteristiche correlate all'identità di gruppo. Per essere chiari, se la caratteristica è statisticamente, causalmente o moralmente irrilevante, ciò potrebbe essere un buon motivo per omittirla (Capitolo 2). Ma cosa succede se la caratteristica è effettivamente rilevante per il risultato? Ad esempio, supponiamo che le persone che contribuiscono a progetti software open source tendano ad essere ingegneri del software migliori. Questo effetto agisce attraverso un percorso causale moralmente rilevante perché i programmati acquisiscono utili competenze di ingegneria del software attraverso la partecipazione open source. Sfortunatamente, molte comunità open source sono ostili e discriminatorie nei confronti delle donne e delle minoranze (forse perché mancano delle strutture organizzative formali che le aziende utilizzano per tenere sotto controllo in una certa misura la discriminazione interpersonale). Riconoscendo ciò, un'azienda di software potrebbe tenerne esplicitamente conto nelle decisioni di assunzione o semplicemente omettere la considerazione dei contributi open source come criterio. Se si adotta la seconda opzione, si ottengono in media assunzioni meno qualificate; svantaggia anche le persone che hanno sfidato la discriminazione per sviluppare le proprie capacità, probabilmente il gruppo più meritevole.

L'omissione di elementi basati su considerazioni statistiche senza una giustificazione morale o causale è estremamente popolare nella pratica perché è semplice da implementare, politicamente appetibile ed evita il rischio legale di un trattamento dispari.

Lotta alla discriminazione interpersonale

Piuttosto che intervenire direttamente sugli output, le organizzazioni possono provare a migliorare il processo decisionale. In molti casi, i discriminatori sono sorprendentemente sinceri riguardo ai loro pregiudizi nei sondaggi e nelle interviste.³²⁴ È forse possibile addestrarli a liberarsi dai loro pregiudizi impliciti o palesi? Questa è l'idea alla base della riduzione dei pregiudizi, spesso chiamata formazione alla diversità.

Ma la formazione sulla diversità funziona? Paluck & Green ha condotto un'analisi approfondita di quasi un migliaio di interventi di questo tipo nel 2009. ⁴⁵⁴ Gli interventi includono la promozione del contatto con membri di diversi gruppi, la ricategorizzazione dell'identità sociale, l'istruzione esplicita, la sensibilizzazione, il targeting delle emozioni, il targeting della coerenza dei valori e dell'autostima, l'apprendimento cooperativo, l'intrattenimento (lettura, media), la discussione e l'influenza dei pari. Sfortunatamente, solo una piccola parte degli studi pubblicati riportano esperimenti sul campo; Paluck e Green sono dubiosi sia sugli studi osservazionali sul campo che sugli esperimenti di laboratorio. Nel complesso, gli esperimenti sul campo non forniscono molto supporto all'efficacia degli interventi sulla diversità. Detto questo, c'erano molti metodi di laboratorio promettenti che non erano ancora stati testati sul campo. Una revisione più recente riassume i progressi della ricerca dal 2007 al 2019.

455

Minimizzare il ruolo del giudizio umano attraverso la formalizzazione

Approcci come la formazione sui pregiudizi impliciti cercano di migliorare il giudizio dei decisori umani, ma alla fine si rimettono a quel giudizio. Al contrario, la formalizzazione mira a frenare il giudizio e la discrezionalità.

La tecnica di formalizzazione più semplice consiste nel nascondere al decisore l'identità del soggetto decisionale (o altre caratteristiche considerate irrilevanti).

Sebbene questa idea risalga all'antichità, in molti ambiti l'adozione della valutazione anonima è un fenomeno recente ed è stata resa più semplice dalla tecnologia.⁴⁵⁶ Due limiti principali di questo approccio sono l'onnipresente disponibilità di proxy e il fatto che in molti casi l'anomimizzazione non è fattibile. contesti come i colloqui di assunzione di persona.⁷ Un approccio più ambizioso è un processo

decisionale basato su regole o statistici che elimina completamente la discrezionalità umana. Ad esempio, l'eliminazione della discrezionalità del prestatore nella sottoscrizione dei prestiti è stata associata a un aumento di quasi il 30% nei tassi di approvazione dei richiedenti appartenenti a minoranze e a basso reddito, aumentando allo stesso tempo l'accuratezza predittiva (del rischio di default).¹ I decisori umani tendono ignorare selettivamente le irregolarità nella storia creditizia dei richiedenti bianchi.⁴⁵⁸

In un certo senso, l'apprendimento automatico può essere visto come una progressione naturale del passaggio dal giudizio umano al processo decisionale basato su regole.

Nell'apprendimento automatico, la scoperta della regola – e non solo la sua applicazione – viene rinviata ai dati e implementata da un sistema automatizzato. Sulla base di ciò, si potrebbe ingenuamente sperare che l'apprendimento automatico sia ancora più efficace nel ridurre al minimo la discriminazione.

Tuttavia, ci sono diverse controargomentazioni. In primo luogo, le affermazioni sulla superiorità delle formule statistiche rispetto al giudizio umano, almeno in alcuni ambiti, sono state messe in dubbio in quanto basate su confronti tra mele e arance perché gli esperti umani non consideravano il loro ruolo come pura previsione. Ad esempio, i giudici che prendono decisioni sulle sentenze possono prendere in considerazione i desideri delle vittime e trattare i giovani come un fattore moralmente a discarico meritevole di clemenza.² In secondo luogo, sono stati riconosciuti tutti i modi in cui l'apprendimento automatico può essere discriminatorio, il che è di natura ovviamente un tema centrale di questo libro. In terzo luogo, ci sono numerosi potenziali inconvenienti, come la perdita di spiegabilità e gli effetti strutturali, che non vengono colti dai confronti uomo-macchina.

Forse la cosa più significativa è che una formalizzazione incompleta può semplicemente spostare l'abuso di discrezionalità altrove. Nel Kentucky, l'introduzione della valutazione del rischio preliminare ha aumentato le disparità razziali per gli imputati con lo stesso rischio previsto. L'effetto sembra essere in parte dovuto alla diversa adozione della valutazione del rischio in contee con diversi dati demografici razziali, e in parte perché anche gli stessi giudici hanno maggiori probabilità di ignorare la decisione raccomandata per gli imputati neri rispetto agli imputati bianchi. ⁴⁵⁹ Stevenson 2018 Assessing, Albright 2019 If In Ontario, gli assistenti sociali dei servizi sociali hanno descritto come manipolano gli input del sistema automatizzato per ottenere i risultati desiderati.⁴⁵⁹⁸ A Los Angeles, gli agenti di polizia hanno utilizzato molte strategie per resistere alla gestione degli algoritmi di polizia predittiva.⁴⁶⁰

L'effetto più pernicioso della formalizzazione come intervento sull'equità è proprio questo

⁷Anche in questi contesti, l'occultamento degli attributi non facilmente deducibili può essere efficace. In effetti, è disapprovato chiedere informazioni sullo stato civile dei candidati durante i colloqui di lavoro e tali domande possono essere trattate come prova di intenti discriminatori.⁴⁵⁷ Gli operatori sociali riferiscono di

farlo per aggirare i limiti e la mancanza di trasparenza del sistema automatizzato per ottenere risultati giusti per i clienti. La difficoltà di distinguere tra abuso di discrezionalità e aggiramento di un sistema eccessivamente rigido illustra ulteriormente la natura a doppio taglio della formalizzazione come intervento di equità.

possono spostare la discrezionalità alle fasi precedenti del processo, rendendo la discriminazione più difficile da mitigare. Gli esempi abbondano. Negli Stati Uniti negli anni '80 le linee guida obbligatorie sulla pena minima per il possesso di droga erano giustificate in parte come un modo per combattere i pregiudizi e l'arbitrarietà dei giudici,⁴⁶¹ ma sono ora ampiamente riconosciute come eccessivamente punitive e strutturalmente razziste. Un modo in cui tali leggi possono codificare la razza è la disparità di condanna di 100 a 1 tra cocaina in polvere e crack, la popolarità delle due forme della stessa droga che differisce in base al reddito e allo status socioeconomico.⁴⁶² Un tipo di esempio molto diverso viene da Google, che ha adottato un processo di reclutamento decentrato e altamente formalizzato al fine di combattere i pregiudizi inconsci e migliorare la qualità delle decisioni.⁴⁶³ Ma i reclutatori hanno sostenuto che questo processo in realtà alimenta la discriminazione razziale perché incorpora una classifica dei college in cui storicamente i college e le università neri non sono affatto classificati.⁴⁶⁴

La causa di ammissione ad Harvard del Capitolo 5 è un altro caso di studio sulla formalizzazione rispetto al processo decisionale olistico. I ricorrenti sottolineano che i criteri di ammissione includono valutazioni soggettive di tratti della personalità quali simpatia, integrità, disponibilità, gentilezza e coraggio. Harvard ha ottenuto in media un punteggio molto più basso per i candidati asiatico-americani su questi tratti rispetto a qualsiasi altro gruppo razziale. Harvard, d'altro canto, sostiene che valutare la "persona nella sua interezza" è importante per identificare coloro che hanno esperienze di vita uniche che potrebbero contribuire alla diversità del campus, e che la considerazione dei tratti soggettivi è una componente necessaria di questa valutazione.

Tutele procedurali

La formazione e la formalizzazione della diversità sono esempi di interventi di equità procedurale. Esistono molte altre tutele procedurali: in particolare, rendere il processo trasparente, fornire spiegazioni sulle decisioni e consentire ai soggetti della decisione di contestare decisioni che potrebbero essere state prese per errore. Come discusso in precedenza, le protezioni procedurali sono più importanti quando è coinvolto l'apprendimento automatico che per altri tipi di sistemi automatizzati.

La legge degli Stati Uniti enfatizza l'equità procedurale rispetto ai risultati. Questa è una delle ragioni della grande popolarità della formazione sulla diversità, nonostante la sua discutibile efficacia.⁴⁶⁵ Quando il decisore è il governo, la concezione giuridica dell'equità è ancora più focalizzata sulla procedura. Ad esempio, non esiste la nozione di impatto disparato nel diritto costituzionale degli Stati Uniti.

Mentre alcuni interventi procedurali, come la formazione sulla diversità, sono stati ampiamente adottati, molti altri rimangono rari nonostante i loro evidenti vantaggi in termini di equità. Ad esempio, pochi datori di lavoro offrono spiegazioni sincere per il rifiuto del lavoro. I decisori che si rivolgono ai sistemi automatizzati spesso cercano di ridurre i costi e potrebbero quindi essere particolarmente riluttanti ad adottare protezioni procedurali. Uno scenario illustrativo da Amazon, che utilizza un sistema automatizzato per gestire i fattori di consegna dei contratti, inclusa la risoluzione del contratto: gli addetti ai lavori hanno riferito che "era più economico fidarsi degli algoritmi che pagare persone per indagare su licenziamenti errati fintanto che i conducenti potevano essere

sostituiti facilmente."⁴⁶⁶ Esistono molti esempi di problemi di equità con i sistemi automatizzati per

cui solo le tutele procedurali possono costituire un rimedio efficace (oltre a demolire del tutto il sistema). Ad esempio, la politica di Google prevede la sospensione degli utenti dall'intera suite di servizi se violano i termini di servizio. Ci sono molte segnalazioni aneddotiche di utenti che hanno perso anni di dati personali e professionali, insistono sul fatto che la decisione di Google è stata presa per errore e che il processo di appello di Google non ha portato a una revisione umana significativa della decisione.

Sensibilizzazione

Il resto degli interventi non riguardano il cambiamento del processo decisionale (o dei risultati). Cambiano invece qualcosa riguardo ai soggetti decisionali o all'ambiente organizzativo.

Uno studio del 2018 condotto da Dynarski, Libassi, Michelmore e Owen ha cercato di affrontare il fenomeno sconcertante secondo cui gli studenti a basso reddito tendono a non frequentare college altamente selettivi, anche quando le loro forti credenziali accademiche li qualificano per l'ammissione e nonostante la disponibilità di aiuti finanziari che potrebbero rendere più economico frequentare un istituto selettivo.⁴⁶⁷ Gli autori hanno progettato un intervento in cui hanno inviato volantini a studenti delle scuole superiori a basso reddito informandoli di una nuova borsa di studio presso l'Università del Michigan, e hanno scoperto che, rispetto a un gruppo di controllo, questi gli studenti avevano più del doppio delle probabilità di fare domanda e di iscriversi all'Università. L'effetto era interamente dovuto agli studenti che altrimenti avrebbero frequentato college meno selettivi o non l'avrebbero frequentato affatto. I target di sensibilizzazione erano studenti altamente qualificati identificati sulla base di punteggi di test standardizzati (ACT e SAT), che consentivano all'università di garantire un aiuto finanziario condizionato all'ammissione. Vale la pena ribadire che si è trattato di un intervento puramente informativo: la borsa di studio era disponibile anche per gli studenti del gruppo di controllo, che hanno ricevuto solo cartoline con le scadenze per le iscrizioni all'Università del Michigan.

Nella misura in cui le disparità sono dovute a gruppi svantaggiati che non conoscono le opportunità, gli interventi informativi dovrebbero ridurre tali disparità, ma questo punto non sembra essere ben studiato. Ad esempio, lo studio del Michigan ha mirato all'intervento sugli studenti a basso reddito, quindi non affronta la questione se informare tutti gli studenti possa colmare il divario di reddito.

Intervenire sui fattori causali

Se comprendiamo i fattori causali che portano alla sottoperformance di alcuni individui o gruppi, possiamo intervenire per mitigarli. Come gli interventi informativi, questo approccio cerca di aiutare tutti gli individui piuttosto che semplicemente minimizzare le disparità.

Questo tipo di intervento è estremamente comune. Alcuni esempi: programmi di formazione professionale per ex detenuti per migliorare il welfare e diminuire le possibilità di recidiva; sforzi per rafforzare l'educazione matematica e scientifica per affrontare una presunta carenza di manodopera di ingegneri (un cosiddetto problema del gasdotto); e essenzialmente tutta la sanità pubblica e l'assistenza sanitaria preventiva. L'uso di studi randomizzati e controllati per identificare e intervenire sulle cause della povertà è stato così influente nell'economia dello sviluppo da portare al Premio Nobel 2019 a Duflo, Banerjee e

Kremer.

In un mercato competitivo, come quello di un datore di lavoro in competizione per i lavoratori, questo intervento potrebbe non ripagare un singolo decisore da un punto di vista economico: le persone in cerca di lavoro che hanno beneficiato dell'intervento possono invece scegliere di unirsi ad altre aziende. Sono stati utilizzati molti approcci per superare questo disallineamento degli incentivi. Le imprese possono agire collettivamente oppure lo Stato può finanziare l' intervento. Se un'impresa è sufficientemente grande, i profitti complessivi potrebbero essere così elevati rispetto al costo dell'intervento che il beneficio reputazionale per l'impresa potrebbe essere sufficiente a giustificarlo.

Modificare l'ambiente organizzativo

Se i decisori hanno molte opportunità di intervenire prima del momento della decisione (ad esempio l'assunzione), hanno anche opportunità di intervenire dopo quel momento per garantire che gli individui realizzino il proprio potenziale. Se un'azienda rileva che pochi dipendenti appartenenti a minoranze hanno successo, potrebbe essere perché il posto di lavoro è ostile e discriminatorio.

In altri casi, alcuni individui o gruppi potrebbero aver bisogno di ulteriori soluzioni per rimediare agli svantaggi del passato o a causa di differenze moralmente irrilevanti.

Alcuni esempi: corsi di recupero per studenti svantaggiati, un gruppo di pari per studenti universitari alla prima esperienza, borse di studio basate sui bisogni, una stanza per madri che allattano sul posto di lavoro e alloggi per disabili.

L'accomodamento non è semplicemente una redistribuzione mascherata: non implica (o non è necessario) una preferenza esplicita per il gruppo svantaggiato. Anche se l'alloggio viene messo a disposizione di tutti, a beneficiarne sarà in via preferenziale il gruppo svantaggiato. Ciò è ovvio nel caso, ad esempio, degli alloggi per disabili. In altri casi questo è meno ovvio, ma non per questo meno vero. Anche se gli aiuti finanziari fossero disponibili per tutti gli studenti di un'università, andrebbero a vantaggio in modo differenziato degli studenti a basso reddito.

Tuttavia, gli effetti reali degli accomodamenti possono essere difficili da prevedere e devono essere attentamente misurati empiricamente. Un esempio degno di nota viene da uno studio che dimostra che gli uomini traggono vantaggio da politiche di blocco dell'orologio neutrali rispetto al genere.⁴⁶⁸ Tali politiche nelle università consentono sia agli uomini che alle donne di aggiungere tempo all'orologio con la nascita di un figlio. Sebbene siano spesso adottati nell'interesse dell'equità, lo studio mostra che aumentano i tassi di permanenza in carica degli uomini e abbassano quelli delle donne; ciò è presumibilmente dovuto al fatto che gli uomini sono in grado di essere più produttivi durante il loro lungo periodo di tempo a causa delle differenze nelle responsabilità di cura dei figli o dell'impatto della nascita stessa. Detto questo, va notato che la politica ha due obiettivi di equità: mitigare l'impatto negativo del parto sulla carriera e ridurre le disparità di genere in tali impatti. Presumibilmente la politica riesce comunque a raggiungere il primo obiettivo anche se fallisce il secondo.

Ecco un chiaro esempio di come le politiche organizzative possono causare il fallimento delle persone e di quanto sia facile porvi rimedio. Nella città di New York si registrano ogni anno circa 300.000 casi di reati di basso livello. Gli imputati sono tenuti a comparire in tribunale (ad eccezione dei reati di minore gravità che possono essere risolti per posta). Se non si presentano, i mandati di arresto vengono emessi automaticamente. Storicamente, un notevole 40% degli imputati non si presenta in tribunale. Le conseguenze negative risultanti dalla mancata apparizione (FTA) sono gravi e distribuite in modo diseguale:

ad esempio, i membri di gruppi soggetti a un eccesso di polizia hanno maggiori probabilità di essere arrestati. Sorprendentemente, uno studio condotto da Fishbane, Ouss e Shah ha rilevato che i tassi di FTA sono diminuiti dal 41% al 26% semplicemente ridisegnando il modulo di citazione per renderlo meno confuso e inviando messaggi di testo agli imputati poco prima della data in tribunale⁴⁶⁹!

Considerazioni conclusive

Abbiamo esaminato sette tipologie generali di interventi sull'equità che le organizzazioni possono implementare. La maggior parte di questi interventi migliorano potenzialmente le opportunità per tutti i soggetti decisionali poiché sono motivati da qualche ingiustizia di fondo piuttosto che limitarsi a mitigare alcune disparità. In effetti, gli interventi che mirano ad affrontare un'ingiustizia di fondo potrebbero talvolta aumentare alcune disparità tra i gruppi – una possibilità che sarebbe moralmente giustificata in base a una nozione non comparativa di equità che richiede di trattare ciascun soggetto come dovrebbe essere trattato.⁴⁷⁰

È interessante concentrarsi sui concetti comparativi di equità perché sono facili da quantificare, ma non dovremmo dimenticare le questioni più profonde. Un ambito in cui ciò sembra essere accaduto è l'assunzione algoritmica. Gli strumenti utilizzati nelle assunzioni algoritmiche utilizzano test di giudizio situazionale, test della personalità e talvolta tecniche molto più dubbie, che coinvolgono sempre più l'apprendimento automatico, per lo screening e la selezione dei candidati. Le aziende adottano tali strumenti per ridurre i costi di assunzione, soprattutto per le posizioni a basso salario in cui il costo di assunzione di un lavoratore attraverso il processo tradizionale può essere considerato significativo in relazione al contributo del lavoratore alle entrate dell'impresa nel corso del periodo di impiego .

Questi strumenti sono problematici per molte ragioni. Pur mirando a formalizzare il processo di assunzione, spesso utilizzano attributi che sono moralmente e causalmente irrilevanti per la prestazione lavorativa. HireVue, ad esempio, in precedenza si affidava alle espressioni facciali e alle intonazioni della voce di una persona come parte della sua valutazione automatizzata. Inoltre non riescono ad avere una visione ampia della discriminazione. Concentrarsi esclusivamente sulla minimizzazione delle disparità nei tassi di assunzione tra i gruppi lascia irrisolto il tipo di ambiente che i dipendenti incontreranno una volta assunti. Se in precedenza si prevedeva che i candidati di determinati gruppi avrebbero ottenuto scarsi risultati in un determinato posto di lavoro, il datore di lavoro dovrebbe sforzarsi di comprendere le ragioni di questa differenza di successo, piuttosto che cercare semplicemente di trovare membri di questi gruppi che potrebbero essere in grado di avere successo in condizioni così sfavorevoli. condizioni inospitali o ostili. Gli interventi che promuovono la parità modificano il processo di selezione, ma preservano lo status quo organizzativo, sostenendo l'idea che i candidati selezionati dovrebbero essere in grado di affrontare queste condizioni sufficientemente bene da essere produttivi quanto i loro pari che non affrontano sfide simili . Altre forme di intervento produttive e potenzialmente meno dannose comprendono la formazione sul posto di lavoro (che potrebbe essere intesa come un modo per intervenire sui fattori causali), un feedback significativo per i candidati respinti (che fornirebbe un certo grado di protezione procedurale, ma anche aiutare a guidare gli investimenti futuri dei candidati nel proprio sviluppo) e un approccio strategico per reperire candidati che le aziende con strumenti più accurati potrebbero ora essere più capaci valutare.

L'attenzione ristretta alle disparità può significare che si tiene poco conto della qualità delle decisioni prese dagli strumenti. Strumenti che semplicemente mancano di validità sollevano una serie di preoccupazioni normative. In particolare, le valutazioni che raggiungono una parità demografica approssimativa ma continuano a soffrire di disparità di accuratezza (chiamata anche validità differenziale) possono predisporre i membri di determinati gruppi al fallimento aspettandosi che siano in grado di ottenere risultati migliori di quelli a cui sarebbero attualmente preparati.⁹

Per ribadire, non siamo favorevoli a considerare i criteri di equità statistica come vincoli, almeno in prima istanza. Questo approccio presuppone che la riallocazione sia l'unico intervento disponibile. Invece, se trattiamo i criteri di equità statistica come diagnostici, è probabile che scopriremo problemi più profondi che richiedono una soluzione.

Sfortunatamente, questi rimedi più profondi sono anche più difficili. Richiedono sia l'inferenza causale che la profondità normativa. Questo è ovviamente il motivo per cui vengono spesso ignorati e le questioni fondamentali rimangono irrisolte.

Un esempio calzante: un articolo del 2021 di Stevenson & Mayson analizza l'equità della custodia cautelare in senso non comparativo.⁴⁷¹ Quanto rischioso deve essere un imputato affinché il beneficio atteso per la sicurezza pubblica giustifichi il danno subito dall'imputato dalla detenzione? Utilizzando l'approccio intelligente di chiedere ai destinatari del sondaggio di scegliere tra essere detenuti o diventare vittime di determinati crimini, gli autori concludono che la custodia cautelare essenzialmente non è mai giustificata.

Il metodo dello studio sarà sicuramente oggetto di dibattito, ma resta il fatto che ci sono stati relativamente pochi tentativi quantitativi e di principio per giustificare le soglie di rischio utilizzate nella custodia cautelare. Ci sono state molte altre richieste per porre fine alla custodia cautelare basate su diversi argomenti morali e legali. Quando tali questioni fondamentali continuano ad essere dibattute, sarebbe estremamente prematuro dichiarare "equo" un sistema di custodia cautelare basato sul rischio perché soddisfa alcuni criteri statistici. criterio.

Note del capitolo

La prima parte del capitolo si ispira fortemente alla sociologia della discriminazione.

Una rassegna sulla discriminazione razziale di Pager e Shepherd è un buon punto di accesso a questa letteratura.³²⁵ Small e pager distillano sei lezioni dalla sociologia della discriminazione.³⁹⁷

I modi complessi in cui opera la discriminazione – circuiti di feedback che sostengono diseguaglianze persistenti, molteplici sistemi di discriminazione interconnessi che insieme strutturano la società – significano che i test quantitativi per la discriminazione discussi nel capitolo precedente inquadrono la questione in modo restrittivo e sono intrinsecamente limitati in ciò che possono rivelare. . Per ulteriori informazioni sui limiti dell'approccio quantitativo, vedere il discorso di Narayanan³⁸⁶ o la discussione nella sezione finale dell'articolo di Lang & Spitzer . [lang2020race]

Il capitolo si rivolge poi alla pratica del machine learning. Per rafforzare l'idea che gli sviluppatori di sistemi di apprendimento automatico devono fare molte scelte che richiedono un giudizio normativo durante tutto il processo di sviluppo, vedere il libro di Jessica Eaglin

⁹Si veda anche la trattazione delle limitazioni dell'indipendenza come criterio di equità nel capitolo 3.

caso di studio sulla previsione del rischio di recidiva.⁴⁷² Passando alla questione di cosa dovrebbero fare diversamente gli esperti di tecnologia nel loro lavoro quotidiano, consigliamo il libro Human-Centered Data Science di Aragon, Guha, Kogan, Muller e Neff [aragon2022human], e quello di Ben Green documento che esorta i data scientist a riconoscere che il loro lavoro è politico.⁴⁷³ La parte finale del capitolo sostiene che la maggior

parte degli interventi sull'equità dovrebbero mirare alla cultura e ai processi organizzativi piuttosto che modificare i criteri decisionali.

Ciò significa che la progettazione di interventi efficaci per l'equità richiede la comprensione delle organizzazioni che intendono adottarli. Questa è una vasta area della sociologia; diamo alcuni campioni. Il classico testo di Michael Lipsky Street Level Bureaucracy discute la complessa relazione tra i singoli decisori e l'agenzia governativa in cui sono inseriti.⁴³⁰ Johnson e Zhang decostruiscono il processo attraverso il quale le burocrazie dei servizi sociali elaborano e implementano le politiche, e sostengono i vantaggi della formalizzazione del processo. processo attraverso un processo decisionale algoritmico.⁷²

Per quanto riguarda i testi su burocrazie o organizzazioni specifiche, Misdemeanorland di Issa Kohler-Hausmann si tuffa nei tribunali penali di grado inferiore della città di New York e mostra che lo scopo e il funzionamento del sistema sono quasi diametralmente diversi da come la maggior parte delle persone e la maggior parte dei libri di testo (incluso questo) concepiscono it.⁴⁷⁴ I libri Pedigree³²⁶ e Inside Graduate Admissions³²⁷ fanno luce rispettivamente sui processi di assunzione presso aziende d'élite e sui processi di ammissione nelle scuole di specializzazione. Uberland di Alex Rosenblat descrive le storie e le condizioni di lavoro degli autisti Uber negli Stati Uniti e in Canada.

Appendice: uno sguardo più approfondito ai fattori strutturali

Analizziamo brevemente due fenomeni che aiutano a spiegare la persistenza della disuguaglianza nel lungo periodo: la segregazione e i circuiti di feedback.

Il ruolo della segregazione

Un fattore strutturale che esacerba tutti i meccanismi di discriminazione di cui abbiamo discusso è la segregazione della società lungo le linee dell'identità di gruppo. La segregazione consente probabilmente la discriminazione interpersonale perché l'aumento del contatto tra i gruppi diminuisce il pregiudizio verso gli outgroup – la controversa ipotesi di contatto.⁴⁷⁵ A livello strutturale, la segregazione sostiene la disuguaglianza

perché le opportunità di un individuo per attività economicamente produttive dipendono dal suo capitale sociale, compresa la casa, la comunità, e ambiente educativo. Un filone della letteratura economica ha costruito modelli matematici e simulazioni per comprendere

come le disuguaglianze di gruppo – in particolare le disuguaglianze razziali – sorgono e persistono indefinitamente anche in assenza di discriminazione interpersonale e nonostante l'assenza di differenze intrinseche tra i gruppi. Nel caso estremo, se immaginiamo due o più gruppi appartenenti ad economie non interagenti che crescono allo stesso ritmo, è intuitivamente chiaro che le differenze possono persistere indefinitamente. Se la segregazione è imperfetta, i divari alla fine si colmano? Questo è sensibile alle ipotesi del modello. A Lundberg

e nel modello di Startz i divari alla fine si colmano, anche se in modo estremamente lento.⁴⁷⁶ Nel modello di Bowles et al., in alcune condizioni ciò non avviene;⁴⁷⁷ una ragione è che il gruppo svantaggiato potrebbe affrontare costi più elevati per l'acquisizione di competenze nel mercato del lavoro a causa

di capitale sociale inferiore.⁴⁷⁸ Negli Stati Uniti, dopo la legislazione sui diritti civili degli anni '60 e '70, la segregazione residenziale per razza è andata diminuendo, anche se lentamente. D'altro canto, la segregazione residenziale in base al reddito sembra essere in aumento.⁴⁷⁹

Il ruolo dei cicli di feedback

Esiste un classico modello economico di circoli viziosi nel contesto di un mercato del lavoro.³²¹ Esistono due gruppi di lavoratori e due tipi di lavori: altamente e poco qualificati, con lavori altamente qualificati che richiedono determinate qualifiche per svolgere in modo efficace. Sotto opportune ipotesi (in particolare, i datori di lavoro non possono osservare perfettamente le qualifiche dei lavoratori prima di assumerli, ma solo dopo aver fornito costose attività di formazione sul posto di lavoro) esiste un equilibrio economico in cui si sostiene il seguente ciclo di feedback:

1. Il datore di lavoro pratica una discriminazione salariale tra i due gruppi.
2. Di conseguenza, il gruppo svantaggiato ottiene rendimenti inferiori sugli investimenti in qualifiche.
3. I lavoratori, ritenuti razionali, rispondono a tale differenziale investendo in modo diverso nell'acquisizione di qualifiche, con un gruppo che acquisisce più qualifiche.
4. Il datore di lavoro – ancora una volta, sotto certi presupposti di razionalità – discrimina il salario a causa della differenza osservata nelle qualifiche.

L'importanza di questo modello è che può spiegare la persistenza della disuguaglianza (e della discriminazione) senza assumere differenze intrinseche tra i gruppi e senza che i datori di lavoro discriminino tra lavoratori ugualmente qualificati. Dovrebbe essere visto come se mostrasse solo la possibilità di tali cicli di feedback. Come ogni modello teorico, affermare che un tale ciclo di feedback spiega alcune disparità effettivamente osservate richiederebbe un'attenta convalida empirica.

9

Set di dati

È diventato un luogo comune sottolineare che i modelli di machine learning sono validi quanto lo sono i dati su cui sono formati. Il vecchio slogan “garbage in, garbage out” si applica senza dubbio alla pratica del machine learning, così come lo slogan correlato “bias in, bias out”. Tuttavia, questi proverbi continuano a sottovalutare, e in qualche modo travisare, l’ importanza dei dati per l’apprendimento automatico.

Non è solo l’output di un algoritmo di apprendimento a soffrire di dati di input scadenti. Un set di dati svolge molte altre funzioni vitali nell’ecosistema del machine learning. Il dataset stesso è parte integrante della formulazione del problema. Implicitamente risolve e rende operativo il problema che i professionisti finiscono per risolvere. I set di dati hanno anche plasmato il percorso di intere comunità scientifiche nella loro capacità di misurare e valutare i progressi, supportare le competizioni e interfacciare tra ricercatori del mondo accademico e professionisti dell’industria.

Se così tanto dipende dai dati nell’apprendimento automatico, potrebbe sorprendere che non esista una risposta semplice alla domanda su cosa rende i dati validi e per quale scopo. La raccolta dei dati per le applicazioni di machine learning non ha seguito alcun quadro teorico consolidato, certamente non riconosciuto a priori.

In questo capitolo esamineremo più da vicino i set di dati più diffusi nel campo dell’apprendimento automatico e i benchmark che supportano. Lo useremo per distinguere i diversi ruoli che i set di dati svolgono in contesti scientifici e ingegneristici. Successivamente esamineremo i danni associati ai dati e discuteremo come mitigarli in base al ruolo del set di dati. Concluderemo con alcune indicazioni generali per migliorare le pratiche relative ai dati.

Limitiamo lo scopo di questo capitolo in alcuni modi importanti. La nostra attenzione si concentrerà in gran parte sui set di dati disponibili al pubblico che supportano scopi di formazione e test nella ricerca e nelle applicazioni di apprendimento automatico. La nostra attenzione esclude ampie fasce di pratiche di raccolta dati industriali, sorveglianza e data mining. Sono inoltre esclusi i dati raccolti intenzionalmente per testare ipotesi scientifiche specifiche, come i dati sperimentali raccolti in uno studio medico.

Un tour dei set di dati in diversi domini

La creazione di set di dati nell'apprendimento automatico non segue un quadro teorico chiaro. I set di dati non vengono raccolti per testare un'ipotesi scientifica specifica. In effetti, vedremo che ci sono molti ruoli diversi che i dati svolgono nell'apprendimento automatico. Di conseguenza, ha senso iniziare esaminando alcuni set di dati influenti provenienti da diversi domini per avere un'idea migliore di cosa sono, cosa ha motivato la loro creazione, come hanno organizzato le comunità e quale impatto hanno avuto.

TIMIT

Il riconoscimento vocale automatico è un problema di apprendimento automatico di notevole interesse commerciale. Le sue radici risalgono all'inizio del XX secolo.⁴⁸⁰

È interessante notare che il riconoscimento vocale presenta anche uno dei set di dati di riferimento più antichi, i dati TIMIT (Texas Instruments/Massachusetts Institute for Technology) . La creazione del set di dati è stata finanziata attraverso un programma DARPA del 1986 sul riconoscimento vocale. A metà degli anni ottanta, l'intelligenza artificiale era nel mezzo di un "inverno dei finanziamenti" in cui molte agenzie governative e industriali erano riluttanti a sponsorizzare la ricerca sull'intelligenza artificiale perché spesso prometteva più di quanto poteva offrire. Il responsabile del programma DARPA Charles Wayne ha proposto che un modo per aggirare questo problema fosse stabilire metodi di valutazione più rigorosi. Wayne ha incaricato il National Institute of Standards and Technology di creare e curare set di dati condivisi per il parlato e ha valutato il successo del suo programma in base alle prestazioni nei compiti di riconoscimento su questi set di dati.

Molti ora attribuiscono al programma di Wayne il merito di aver dato il via a una rivoluzione del progresso nel riconoscimento vocale.⁴⁸¹ Secondo Kenneth Ward Church,

Ha consentito l'avvio dei finanziamenti perché il progetto era a prova di glamour e inganno, e di continuare perché i finanziatori potevano misurare i progressi nel tempo. L'idea di Wayne facilita la produzione di grafici che aiutano a vendere il programma di ricerca a potenziali sponsor. Un vantaggio meno evidente dell'idea di Wayne è che consente l'arrampicata in collina. I ricercatori che inizialmente si erano opposti a sottoporsi ai test due volte l'anno hanno iniziato a valutarsi ogni ora.

Un primo prototipo del dataset TIMIT fu pubblicato nel dicembre del 1988 su CD-ROM. Nell'ottobre 1990 seguì una versione migliorata. TIMIT presentava già la suddivisione formazione/test tipica dei moderni benchmark di apprendimento automatico.

Sappiamo molto sulla creazione dei dati grazie alla loro documentazione approfondita.⁴⁸²

TIMIT prevede un totale di circa 5 ore di parlato, composte da 6300 enunciati, nello specifico 10 frasi pronunciate da ciascuno dei 630 relatori. Le frasi sono state tratte da un corpus di 2342 frasi come la seguente.

Ha tenuto il tuo abito scuro nell'acqua unta tutto l'anno. (sa1)
Non chiedermi di portare uno straccio unto come quello. (sa2)

Questo è stato facile per noi. (sx3)

Jane potrebbe guadagnare di più lavorando sodo. (sx4)

Lei è più magra di me. (sx5)

Il sole splendente brilla sull'oceano. (sx6)

Niente è tanto offensivo quanto l'innocenza. (sx7)

La documentazione TIMIT distingue tra le 8 principali regioni dialettali dei Stati Uniti, documentati come New England, Northern, North Midland, South Midland, Southern, New York City, Western, Army Brat (spostato). Dei relatori, il 70% sono maschi e il 30% sono femmine. Tutti i madrelingua dell'inglese americano, i soggetti all'epoca erano principalmente dipendenti della Texas Instruments. Molti di loro lo erano nuovo nell'area di Dallas dove lavoravano.

Le informazioni razziali sono state fornite con la distribuzione dei dati e codificate come "bianco", "nero", "indiano americano", "ispano-americano", "orientale" e "Sconosciuto". Dei 630 parlanti, 578 sono stati identificati come bianchi, 26 come neri, 2 come Indiani d'America, 2 ispano-americani, 3 orientali e 17 sconosciuti.

Tabella 9.1: Informazioni demografiche sui parlanti TIMIT

ehm

	Uomini	Donne	Totale (%)
Bianco	402 176	578 (91,7%)	
Nero 15	11	26 (4,1%)	
Indiano americano 2	0	2 (0,3%)	
Ispano-americano 2	0	2 (0,3%)	
Orientale 3	0	3 (0,5%)	
Sconosciuto	12	5	17 (2,6%)

La documentazione rileva:

Oltre a questi 630 altoparlanti, un piccolo numero di altoparlanti con accenti stranieri o altre anomalie linguistiche e/o uditive estreme sono stati registrati come soggetti "ausiliari", ma non sono inclusi nel CD-ROM.

Non sorprende che i primi modelli di riconoscimento vocale abbiano avuto un ruolo significativo pregiudizi demografici e razziali nelle loro prestazioni.

Oggi, diverse grandi aziende, tra cui Amazon, Apple, Google e Microsoft , utilizzano modelli di riconoscimento vocale in una varietà di prodotti, dai telefoni cellulari ai telefoni cellulari. app agli assistenti vocali. Non esiste più un importante benchmark aperto che lo farebbe supportare modelli formativi competitivi con le controparti industriali. I processi di riconoscimento vocale industriale sono generalmente complessi e utilizzano dati proprietari fonti di cui non sappiamo molto. Tuttavia, il riconoscimento vocale di oggi

i sistemi continuano a mostrare disparità di prestazione lungo le linee razziali.485

Repository per l'apprendimento automatico dell'UCI

L'UCI Machine Learning Repository ospita attualmente più di 500 set di dati, principalmente per diverse attività di classificazione e regressione. La maggior parte dei set di dati sono relativamente piccoli e comprendono poche centinaia o poche migliaia di istanze. La maggior parte sono set di dati tabulari strutturati con una manciata o poche decine di attributi.

L'UCI Machine Learning Repository ha contribuito all'adozione del paradigma del test del treno nell'apprendimento automatico alla fine degli anni '80. Pat Langley ricorda:

Il movimento sperimentale fu aiutato da un altro sviluppo. David Aha, allora studente di dottorato presso l'UCI, iniziò a raccogliere set di dati da utilizzare negli studi empirici sull'apprendimento automatico. Questo è cresciuto nell'UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/>), che ha reso disponibile alla comunità tramite FTP nel 1987. Questo è stato rapidamente adottato da molti ricercatori perché era facile da usare e perché ha permesso loro di confrontare i loro risultati con risultati precedenti sugli stessi compiti.⁴⁸⁶

Il set di dati più popolare nel repository è l'Iris Data Set contenente misurazioni tassonomiche di 150 fiori di iris, 50 per ciascuna delle 3 specie. Il compito è classificare le specie in base alle misurazioni.

A partire da ottobre 2020, il secondo set di dati più popolare nell'archivio UCI è il set di dati per adulti. Estratto dal database del censimento del 1994, presenta quasi 50.000 esempi che descrivono individui negli Stati Uniti, ciascuno con 14 attributi. Il compito è classificare se un individuo guadagna più di 50.000 dollari USA o meno. Il set di dati per adulti rimane popolare nella comunità dell'equità algoritmica, soprattutto perché è uno dei pochi set di dati disponibili al pubblico che presenta informazioni demografiche tra cui il genere (codificato in binario come maschio/femmina), nonché la razza (codificata come Amer-Indian-Eskimo , Asian-Pac-Islander, Nero, Altro e Bianco).

Sfortunatamente, i dati presentano alcune idiosincrasie che li rendono tutt'altro che ideali per comprendere i pregiudizi nei modelli di apprendimento automatico. A causa dell'età dei dati e del limite di reddito di \$ 50.000, quasi tutti i casi etichettati come neri sono al di sotto del limite, così come quasi tutti i casi etichettati come donne. In effetti, un modello di regressione logistica standard addestrato sui dati raggiunge complessivamente una precisione di circa l'85%, mentre lo stesso modello raggiunge una precisione del 91% sulle istanze nere e quasi il 93% sulle istanze femminili. Allo stesso modo, le curve ROC per gli ultimi due gruppi racchiudono in realtà un'area maggiore rispetto alla curva ROC per i casi maschili. Si tratta di una situazione atipica: più spesso, i modelli di machine learning hanno prestazioni peggiori su gruppi storicamente svantaggiati.

MNIST

Il set di dati MNIST contiene immagini di cifre scritte a mano. La sua versione più comune ha 60.000 immagini di training e 10.000 immagini di prova, ciascuna con 28x28 pixel in bianco e nero.

MNIST è stato creato dai ricercatori Burges, Cortes e Lecun da un precedente set di dati rilasciato dal National Institute of Standards and Technology (NIST).



Figura 9.1: Un campione di cifre MNIST

Il set di dati è stato introdotto in un documento di ricerca nel 1998 per mostrare l'uso di metodi di deep learning basati su gradienti per attività di riconoscimento dei documenti.⁴⁸⁷ Da allora, citato più di 30.000 volte, MNIST è diventato un punto di riferimento molto influente nella comunità della visione artificiale. Due decenni dopo, i ricercatori continuano a utilizzare attivamente i dati.

I dati originali del NIST avevano la proprietà che i dati sulla formazione e sui test provenivano da due popolazioni diverse. Il primo presentava la calligrafia di duemila dipendenti dell'American Census Bureau, mentre il secondo proveniva da cinquecento studenti delle scuole superiori americane.⁴⁸⁸ I creatori di MNIST hanno rimescolato queste due fonti di dati e le hanno divise in set di formazione e test. Inoltre, hanno ridimensionato e centrato le cifre. L'esatta procedura per derivare il MNIST dal NIST è andata perduta, ma è stata recentemente ricostruita confrontando le immagini di entrambe le fonti di dati.⁴⁸⁹

Il set di test MNIST originale aveva le stesse dimensioni del set di addestramento, ma il set di test più piccolo è diventato standard nell'uso della ricerca. Le 50.000 cifre del set di test originale che non sono state inserite nel set di test più piccolo sono state successivamente identificate e soprannominate le cifre perse.⁴⁸⁹

Fin dall'inizio, il MNIST doveva essere un punto di riferimento utilizzato per confrontare i punti di forza di diversi metodi. Per diversi anni, LeCun ha mantenuto una classifica informale su un sito web personale che elencava i numeri con la migliore precisione ottenuti da diversi algoritmi di apprendimento su MNIST.

Tabella 9.2: un'istantanea della classifica MNIST originale
dal 2 febbraio 1999. Fonte: Archivio Internet (recuperato:
4 dicembre 2020)

Metodo	Errore del test (%)
classificatore lineare (NN a 1 strato)	12.0
classificatore lineare (NN a 1 strato) [raddrizzamento]	8.4
classificatore lineare a coppie	7.6
K-vicini più vicini, euclidei	5.0
K-vicini più vicini, euclidei, raddrizzati	2.4
40 PCA + classificatore quadratico	3.3
1000 RBF + classificatore lineare	3.6
K-NN, distanza tangente, 16x16	1.1
Polinomio SVM grado 4	1.1
Polinomio ridotto Set SVM deg 5	1.0
Virtual SVM deg 9 poly [distorsioni] 2 strati NN, 300 unità nascoste 2 strati NN, 300	0,8
HU, [distorsioni] 2 strati NN, 300 HU, [deskewing] 2 strati NN, 1000 unità	4.7
nascoste 2 strati NN , 1000 HU,	3.6
[distorsioni] 3 strati NN, 300+100 unità	1.6
nascoste 3 strati NN, 300+100 HU	4.5
[distorsioni] 3 strati NN, 500+150 unità nascoste	3.8
3 strati NN, 500+150 HU [distorsioni]	3.05
	2.5
	2,95
	2.45
LeNet-1 [con ingresso 16x16]	1.7
LeNet-4	1.1
LeNet-4 con K-NN invece dell'ultimo strato	1.1
LeNet-4 con apprendimento locale invece di Il	1.1
LeNet-5, [nessuna distorsione]	0,95
LeNet-5, [enormi distorsioni]	0,85
LeNet-5, [distorsioni]	0,8
LeNet-4 potenziato, [distorsioni]	0,7

Nella sua veste di punto di riferimento, è diventato una vetrina per il kernel emergente metodi dei primi anni 2000 che hanno raggiunto temporaneamente le massime prestazioni su MNIST.⁴⁹⁰ Oggi non è difficile ottenere un errore di classificazione inferiore allo 0,5% con un errore di classificazione ampio gamma di architetture di reti neurali convoluzionali. I migliori modelli si classificano tutti tranne alcuni casi di test patologici correttamente. Di conseguenza, MNIST è ampiamente considerato troppo facile per gli odierni compiti di ricerca.

MNIST non è stato il primo set di dati di cifre scritte a mano utilizzato per l'apprendimento automatico ricerca. In precedenza, il servizio postale degli Stati Uniti (USPS) aveva rilasciato un set di dati di 9298 immagini (7291 per la formazione e 2007 per i test). I dati USPS erano in realtà un bel po' più difficile da classificare rispetto a MNIST. Una frazione non trascurabile dell'aspetto delle cifre USPS

irriconoscibili per gli esseri umani,⁴⁹¹ mentre gli esseri umani riconoscono essenzialmente tutte le cifre in MNIST.

ImageNet

ImageNet è un ampio archivio di immagini etichettate che ha avuto una grande influenza nella ricerca sulla visione artificiale negli ultimi dieci anni. Le etichette delle immagini corrispondono ai nomi del database lessicale WordNet della lingua inglese.⁴⁹² WordNet raggruppa i nomi in sinonimi cognitivi, chiamati synset. Le parole automobile e automobile, ad esempio, rientrerebbero nello stesso synset. Oltre a queste categorie WordNet fornisce una struttura ad albero gerarchico secondo una relazione super-subordinata tra synset. Il synset per sedia, ad esempio, è figlio del synset per mobili nella gerarchia wordnet. WordNet esisteva prima di ImageNet e in parte ha ispirato la creazione di Imagenet.

La versione iniziale di ImageNet includeva circa 5000 categorie di immagini, ciascuna corrispondente a un synset in WordNet. Queste categorie ImageNet contenevano in media circa 600 immagini per categoria.⁴⁹³ ImageNet è cresciuto nel tempo e la sua versione dell'autunno 2011 ha raggiunto circa 32.000 categorie.

La costruzione di ImageNet ha richiesto due passaggi essenziali: recuperare le immagini candidate per ciascun synset ed etichettare le immagini recuperate. Per questo primo passo sono stati utilizzati motori di ricerca online e piattaforme di condivisione di foto con un'interfaccia di ricerca, in particolare Flickr. Le immagini dei candidati sono state prese dai risultati della ricerca di immagini associati ai nomi synset per ciascuna categoria.

Per la seconda fase di etichettatura, i creatori di ImageNet si sono rivolti alla piattaforma Mechanical Turk (MTurk) di Amazon . MTurk è un mercato del lavoro online che consente a individui e aziende di assumere lavoratori su richiesta per svolgere compiti semplici. In questo caso, ai lavoratori di MTurk sono state presentate le immagini candidate e hanno dovuto decidere se l'immagine candidata era effettivamente o meno un'immagine corrispondente alla categoria a cui era presumibilmente associata.

È importante distinguere tra questo database ImageNet e un popolare benchmark e concorso di machine learning, chiamato ImageNet Large Scale Visual Recognition Challenge (ILSVRC), che ne è derivato.⁴⁹⁴ Il concorso è stato organizzato ogni anno dal 2010 al 2017, raggiungendo una significativa notorietà in sia l'industria che il mondo accademico, soprattutto come punto di riferimento per i modelli emergenti di deep learning.

Quando i professionisti dell'apprendimento automatico dicono "ImageNet", in genere si riferiscono ai dati utilizzati per l'attività di classificazione delle immagini nel benchmark ILSVRC del 2012. La competizione prevedeva altri compiti, come il riconoscimento degli oggetti, ma la classificazione delle immagini è diventata il compito più popolare per il set di dati. Espressioni come "un modello addestrato su ImageNet" si riferiscono tipicamente all'addestramento di un modello di classificazione delle immagini sul set di dati di riferimento del 2012.

Un'altra pratica comune che coinvolge i dati ILSVRC è la pre-formazione. Spesso un professionista ha in mente un problema di classificazione specifico il cui insieme di etichette differisce dalle 1000 classi presenti nei dati. È comunque possibile utilizzare i dati per creare funzionalità utili che possono poi essere utilizzate nel problema di classificazione del target.

Il punto in cui ILSVRC entra nelle applicazioni del mondo reale è spesso per supportare la pre-formazione.

Questo uso colloquiale della parola ImageNet può creare confusione, anche perché il set di dati ILSVRC-2012 differisce in modo significativo dal database più ampio. Include solo un sottoinsieme di 1000 categorie. Inoltre, queste categorie sono un sottoinsieme piuttosto distorto della più ampia gerarchia ImageNet. Ad esempio, di queste 1000 categorie solo tre si trovano nel ramo persona della gerarchia di WordNet, in particolare sposo, giocatore di baseball e subacqueo. Tuttavia, più di 100 categorie su 1000 corrispondono a razze canine diverse. Il numero è 118, per l'esattezza, senza contare lupi, volpi e licaoni che sono presenti anche tra le 1000 categ-

Ciò che ha motivato la scelta esatta di queste 1000 categorie non è del tutto chiaro. L'apparente inclinazione canina, tuttavia, non è nemmeno solo una stranezza. All'epoca, nella comunità della visione artificiale c'era interesse a fare progressi nella previsione con molte classi, alcune delle quali sono molto simili. Ciò riflette un modello più ampio nella comunità del machine learning. La creazione di set di dati è spesso guidata da un senso intuitivo di quali siano le sfide tecniche per il settore.

Nel caso di ImageNet, un'altra considerazione importante è stata la scala, sia in termini di numero di immagini che di numero di classi.

Le annotazioni e le etichettature su larga scala inserite in Imagenet rientrano in una categoria di lavoro che Gray e Suri chiamano lavoro fantasma nel loro libro omonimo.⁴⁹⁵ Sottolineano :

I lavoratori di MTurk sono gli eroi non celebrati della rivoluzione dell'intelligenza artificiale.

In effetti, ImageNet è stata etichettata da circa 49.000 lavoratori turchi provenienti da 167 paesi prova nel corso di più anni.

Il Premio Netflix

Il Premio Netflix è stato uno dei concorsi di machine learning più famosi.

A partire dal 2 ottobre 2006, il concorso è durato quasi tre anni e si è concluso con un primo premio di 1 milione di dollari, annunciato il 18 settembre 2009. Nel corso degli anni, il concorso ha visto 44.014 iscrizioni da 5.169 squadre.

I dati di addestramento di Netflix contenevano circa 100 milioni di valutazioni di film da quasi 500mila abbonati Netflix su una serie di 17770 film. Ogni punto dati corrisponde a una tupla <utente, film, data di valutazione, valutazione>. Con una dimensione di circa 650 megabyte, il set di dati era abbastanza piccolo da stare su un CD-ROM, ma abbastanza grande da rappresentare una sfida in quel momento.

I dati Netflix possono essere pensati come una matrice con $n = 480189$ righe e $m = 17770$ colonne. Ogni riga corrisponde a un abbonato Netflix e ogni colonna a un film. Le uniche voci presenti nella matrice sono quelle per le quali un dato abbonato ha valutato un dato film con valutazione in {1, 2, 3, 4, 5}. Tutte le altre voci, ovvero la stragrande maggioranza, mancano. L'obiettivo dei partecipanti era quello di prevedere le voci mancanti della matrice, un problema noto come completamento della matrice o, più in generale, filtraggio collaborativo. In effetti, la sfida di Netflix ha contribuito così tanto a rendere popolare questo problema che a volte viene chiamato il problema Netflix.

L'idea è che se potessimo prevedere le voci mancanti, saremmo in grado di consigliare di conseguenza film mai visti agli utenti.

I dati di riserva che Netflix ha tenuto segreti consistevano in circa tre milioni di ascolti. La metà di essi è stata utilizzata per calcolare una classifica corrente durante la competizione. L'altra metà ha determinato il vincitore finale.

La concorrenza di Netflix è stata estremamente influente. Non solo attirò una partecipazione significativa, ma alimentò anche molto interesse accademico nei confronti del filtraggio collaborativo negli anni a venire. Inoltre, ha reso popolare il formato del concorso come un modo interessante per le aziende di interagire con la comunità del machine learning. Una startup chiamata Kaggle, fondata nell'aprile 2010, ha organizzato centinaia di concorsi di machine learning per varie aziende e organizzazioni prima della sua acquisizione da parte di Google nel 2017.

Ma la concorrenza di Netflix è diventata famigerata per un altro motivo. Sebbene Netflix avesse sostituito i nomi utente con numeri pseudonimi, i ricercatori Narayanan e Shmatikov sono riusciti a identificare nuovamente alcuni degli abbonati Netflix le cui valutazioni dei film erano presenti nel set di dati⁴⁹⁶ collegando tali valutazioni con le valutazioni dei film pubblicamente disponibili su IMDB, un database di film online. Alcuni abbonati Netflix avevano anche valutato pubblicamente una serie di film sovrapposti su IMDB con la loro vera identità. Nella letteratura sulla privacy, questo è chiamato attacco di collegamento ed è uno dei modi in cui dati apparentemente anonimizzati possono

essere resi anonimi.⁴⁹⁷ Ciò che seguì furono molteplici azioni legali collettive contro Netflix, nonché un'indagine della Federal Trade Commission sulla privacy. preoccupazioni. Di conseguenza, Netflix ha annullato i piani per un secondo concorso, annunciato il 6 agosto 2009.

Ad oggi, le preoccupazioni sulla privacy rappresentano un ostacolo legittimo al rilascio pubblico dei dati e alla creazione di set di dati. Le tecniche di deanonymizzazione sono mature ed efficienti. Probabilmente non esiste un algoritmo in grado di prendere un set di dati e fornire una rigorosa garanzia di privacy a tutti i partecipanti, pur essendo utile per tutte le analisi e scopi di apprendimento automatico. Dwork e Roth la chiamano la Legge Fondamentale del Recupero delle Informazioni: "risposte eccessivamente accurate a troppe domande distruggeranno la privacy in modo spettacolare".⁴⁹⁸

Ruoli giocati dai set di dati

Nella ricerca e nell'ingegneria sull'apprendimento automatico, i set di dati svolgono un insieme di ruoli diversi e più importanti rispetto alla maggior parte degli altri campi. Ne abbiamo menzionati diversi in precedenza, ma ora li esaminiamo più nel dettaglio. Comprenderli è fondamentale per capire quali aspetti tecnici e culturali dei benchmark sono essenziali, come si verificano i danni e come mitigarli.

Una fonte di dati reali

Edgar Anderson era un botanico e orticoltore che trascorse gran parte degli anni '20 e '30 raccogliendo e analizzando dati sugli iris per studiare questioni biologiche e tassonomiche. Il set di dati Iris menzionato nel repository di machine learning dell'UCI

sopra c'è il risultato del lavoro di Anderson - o una piccola parte di esso, poiché la maggior parte delle osservazioni nel set di dati provengono da un singolo giorno di lavoro sul campo. Il set di dati contiene 50 osservazioni ciascuna di 3 piante di iris; il compito è quello di distinguere la specie in base a 4 attributi fisici (lunghezza e larghezza dei sepali; lunghezza e larghezza dei petali).

La maggior parte delle decine di migliaia di ricercatori che hanno utilizzato questo set di dati non sono interessati alla tassonomia, per non parlare degli iris. Per cosa, allora, stanno utilizzando il set di dati?

Sebbene i dati siano stati raccolti da Anderson, in realtà furono pubblicati nell'articolo "L'uso di misurazioni multiple nei problemi tassonomici" di Ronald Fisher, fondatore della statistica moderna nonché eugenetista.⁴⁹⁹ Il collegamento con l'eugenetica non è casuale: altre figure centrali nello sviluppo della statistica moderna, come Francis Galton e Karl Pearson, erano algoeugeneticici.⁵⁰⁰,⁵⁰¹ Fisher fu collaboratore di Anderson. Sebbene Fisher avesse un certo interesse per la tassonomia, era principalmente interessato a utilizzare i dati per sviluppare tecniche statistiche (con un occhio verso le applicazioni per l'eugenetica). Nell'articolo del 1936, Fisher introduce l'analisi discriminante lineare (LDA) e mostra che funziona bene in questo compito.

Il motivo per cui il set di dati Iris si è rivelato una buona applicazione di LDA è che esiste una proiezione lineare delle quattro caratteristiche che sembra risultare in una miscela di gaussiane (una per ciascuna delle tre specie) e le medie delle tre distribuzioni sono relativamente distanti; una delle specie è infatti perfettamente separabile dalle altre due. Ogni algoritmo di apprendimento fa implicitamente delle ipotesi sul processo di generazione dei dati: senza ipotesi, non esiste alcuna base per fare previsioni su punti invisibili.⁵⁰² Se potessimo descrivere perfettamente matematicamente il processo di generazione dei dati dietro le caratteristiche fisiche degli iris (o di qualsiasi altra popolazione), non avremmo bisogno di un set di dati: potremmo calcolare matematicamente il rendimento di un algoritmo. In pratica, per fenomeni complessi, raramente esistono descrizioni matematiche così perfette. Comunità diverse attribuiscono un valore diverso al tentativo di scoprire il vero processo di generazione dei dati. L'apprendimento automatico attribuisce relativamente poca importanza a questo obiettivo.⁵⁰³ In definitiva, l'utilità di un algoritmo di apprendimento viene stabilita testandolo su set di dati reali.

La dipendenza dai set di dati di riferimento come fonte di dati reali è stato uno sviluppo graduale nella ricerca sull'apprendimento automatico. Ad esempio, gli esperimenti sul percepitrone di Rosenblatt negli anni '50 utilizzarono due stimoli artificiali (i caratteri E e X), con numerose varianti di ciascuno creati dalla rotazione e da altre trasformazioni.⁵⁰⁴ L'input controllato era considerato utile per comprendere il comportamento del sistema. In un articolo del 1988, Pat Langley sostiene un approccio ibrido, sottolineando che "gli esperimenti riusciti su un certo numero di domini naturali diversi forniscono prova di generalità", ma evidenziando anche l'uso di dati artificiali per una migliore comprensione.⁵⁰⁵ Soprattutto dopo la creazione dell'archivio dell'UCI, in questo periodo è diventato comune valutare nuovi algoritmi su set di dati di benchmark ampiamente utilizzati come un modo per stabilire che il ricercatore non sta "imbrogliando" selezionando input artificiosi.

Per riassumere, quando un ricercatore cerca di presentare prove dell'utilità di un'innovazione algoritmica, l'uso di set di dati reali anziché dati artificiali garantisce che il ricercatore non abbia inventato dati per adattarli all'algoritmo. Inoltre, l'uso di

importanti set di dati di riferimento scongiurano lo scetticismo sul fatto che il ricercatore possa aver scelto un set di dati con proprietà specifiche che rendono l'algoritmo efficace. Infine, l'uso di più set di dati di benchmark provenienti da domini diversi suggerisce che l'algoritmo è altamente generale.

Perversamente, l'ignoranza del dominio viene trattata quasi come una virtù piuttosto che come uno svantaggio. Ad esempio, i ricercatori che raggiungono prestazioni all'avanguardia su (diciamo) La traduzione dal cinese all'inglese potrebbe indicare che nessuno di loro parla cinese.

Il sottotesto è che non avrebbero potuto scegliere consapevolmente o inconsapevolmente un modello che funziona bene solo quando la lingua di partenza è linguisticamente simile al cinese.

Un catalizzatore e una misura del progresso specifico del settore

Le innovazioni algoritmiche altamente portabili tra domini, sebbene importanti, sono rare. Gran parte dei progressi nel campo dell'apprendimento automatico sono invece adattati a settori e problemi specifici. Il modo più comune per dimostrare tali progressi è dimostrare che l'innovazione in questione può essere utilizzata per raggiungere prestazioni "all'avanguardia" su un set di dati di riferimento per tale compito.

L'idea che i set di dati stimolino l'innovazione algoritmica ha qualche spiegazione. Ad esempio, il Premio Netflix è comunemente considerato responsabile della scoperta dell'efficacia della fattorizzazione a matrice nei sistemi di raccomandazione (spesso attribuita a Simon Funk, un concorrente pseudonimo⁵⁰⁶). Tuttavia, la tecnica era stata proposta nel contesto della raccomandazione di film già nel 1998⁵⁰⁷ e per la 1990 avesse ricerca 508 Tuttavia, non era precedentemente evidente che già nel sovraperformato i metodi basati sul quartiere e che potesse scoprire fattori latenti significativi. La chiarezza della classifica Netflix e la credibilità del set di dati hanno contribuito a stabilire l'importanza della fattorizzazione della matrice.⁵⁰⁹

In un certo senso separatamente dal ruolo di stimolare l'innovazione algoritmica, i set di dati di benchmark offrono anche un modo conveniente per misurarne i risultati (da cui il termine benchmark). La progressione dell'accuratezza all'avanguardia su un set di dati e un'attività di riferimento può essere un indicatore utile. Una curva di precisione relativamente piatta nel tempo può indicare che il progresso si è bloccato, mentre un salto discontinuo può indicare una svolta. Raggiungere un tasso di errore vicino allo zero o almeno inferiore all'"errore umano" per i compiti di percezione è spesso considerato un segno che il compito è "risolto" e che è ora che la comunità passi a una sfida più difficile.

Sebbene queste siano euristiche allettanti, ci sono anche delle insidie. In particolare, un'affermazione come "la precisione allo stato dell'arte per la classificazione delle immagini è del 95%" non è un'affermazione scientificamente significativa a cui può essere assegnato un valore di verità, perché il numero è altamente sensibile alla distribuzione dei dati.

Un esempio notevole di questo fenomeno viene da un articolo di Recht, Roelofs, Schmidt e Shankar. Hanno ricreato attentamente nuovi set di test per i parametri di classificazione CIFAR-10 e ImageNet secondo la stessa procedura dei set di test originali.⁵¹⁰ Hanno poi preso un'ampia raccolta di modelli rappresentativi proposti nel corso degli anni e li hanno valutati tutti nel nuovo test insieme. Tutti i modelli hanno subito un calo significativo delle prestazioni sul nuovo set di test, corrispondente a circa 5 anni di progressi nella classificazione delle immagini. Hanno scoperto che questo è perché

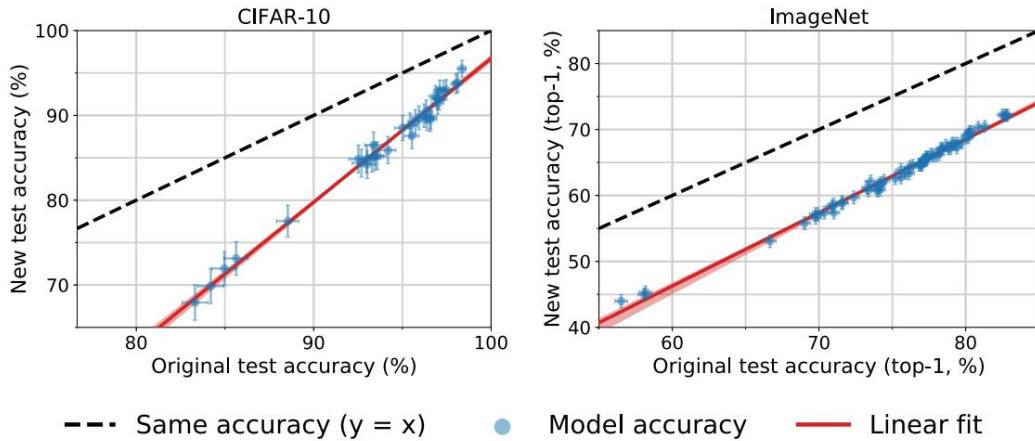


Figura 9.2: Accuratezza del modello sui set di test originali rispetto ai nuovi set di test per CIFAR-10 e ImageNet. Ciascun punto dati corrisponde a un modello in un banco di prova di modelli rappresentativi (mostrati con intervalli di confidenza Clopper-Pearson al 95%). I grafici rivelano due fenomeni principali: (i) C'è generalmente un calo significativo nell'accuratezza dal set di test originale a quello nuovo. (ii) L'accuratezza del modello segue da vicino una funzione lineare, il che significa che i modelli che funzionano bene sul vecchio set di test tendono a funzionare bene anche sul nuovo set di test. La regione ombreggiata stretta è una regione di confidenza del 95% per l'adattamento lineare.

il nuovo set di test rappresenta una distribuzione leggermente diversa. Ciò nonostante gli attenti sforzi dei ricercatori per replicare la procedura di raccolta dei dati; dovremmo aspettarci che i set di test creati da procedure diverse diano luogo a differenze di prestazioni molto maggiori.

Gli stessi grafici forniscono anche un'illustrazione sorprendente del motivo per cui i set di dati benchmark sono una necessità pratica per il confronto delle prestazioni nell'apprendimento automatico. Consideriamo un ipotetico approccio alternativo analogo alla norma in molti altri rami della scienza: un ricercatore che valuta un'affermazione (algoritmo) descrive in dettaglio la sua procedura per campionare i dati; altri ricercatori che lavorano sullo stesso problema campionano i propri set di dati in base alla procedura pubblicata. Si verifica un certo riutilizzo dei set di dati, ma non vi è alcuna standardizzazione. I grafici mostrano che anche sforzi estremamente accurati per campionare un nuovo set di dati dalla stessa distribuzione sposterebbero la distribuzione in modo sufficiente da rendere senza speranza il confronto delle prestazioni.

In altre parole, i dati sull'accuratezza riportati dai set di dati di riferimento non costituiscono conoscenza scientifica generalizzabile, perché non hanno validità esterna oltre il set di dati specifico. Mentre il Recht et al. L'articolo è limitato alla classificazione delle immagini, sembra scientificamente prudente presumere una mancanza di validità esterna anche per altri compiti di apprendimento automatico, a meno che non vi sia prova contraria. Tuttavia, i due grafici qui sopra suggeriscono un diverso tipo di conoscenza che sembra trasferirsi quasi perfettamente al nuovo set di test: la prestazione relativa dei modelli. In effetti, un altro articolo ha dimostrato che la performance relativa è stabile su molti set di dati in una gamma molto più ampia di cambiamenti distributivi, con forti

correlazioni tra prestazioni interne ed esterne al dominio.⁵¹¹

La prestazione relativa dei modelli per un determinato compito è un tipo molto utile di conoscenza orientata al professionista che può essere acquisita dalle classifiche di riferimento. Una domanda che spesso i professionisti si trovano ad affrontare è: "quale classe di modelli dovrebbero utilizzare per [un dato compito] e come dovrebbero ottimizzarlo"? Un set di dati di riferimento (insieme alla definizione del compito associato) può essere visto come un proxy per rispondere a questa domanda in un contesto limitato, analogo agli studi di laboratorio in altri rami della scienza. La speranza è che gli algoritmi (e le classi o le architetture del modello) identificati come lo stato dell'arte sulla base della valutazione benchmark siano anche quelli che saranno efficaci sul set di test del professionista. In altre parole, i professionisti possono affidare il laborioso compito di selezione del modello alla classifica dei benchmark.

Per essere chiari, questa è una semplificazione eccessiva. I professionisti hanno molte preoccupazioni oltre all'accuratezza, come il costo computazionale (sia della formazione che della previsione), l'interpretabilità e, sempre più, l'equità e il costo ambientale.

Pertanto, la prestazione del benchmark è utile per i professionisti ma non è l'unica considerazione per la selezione del modello.

Possiamo immaginare uno spettro di quanto il nuovo set di test sia simile al set di benchmark. Da un lato, se il nuovo set di test è veramente un nuovo campione della stessa identica distribuzione, allora la classificazione delle classi del modello dovrebbe essere la stessa per i due set. All'estremo opposto, le distribuzioni possono essere così diverse da costituire compiti essenzialmente diversi, così che la performance di uno non è una guida utile per la performance dell'altro. Tra questi estremi c'è una grande zona grigia che non è ben compresa, e che attualmente è più arte che scienza.

La mancanza di chiarezza su quanto possiamo generalizzare a partire da uno o pochi parametri di riferimento è associata a controversie ben note. Ad esempio, le macchine a vettori di supporto erano competitive con le reti neurali su parametri di riferimento della generazione precedente come il riconoscimento delle cifre del NIST,⁵¹² e questa è stata una delle ragioni per cui l'interesse per le reti neurali è diminuito negli anni '90. La netta superiorità delle reti neurali rispetto ai benchmark più recenti come ImageNet è stata riconosciuta solo tardivamente.¹

Una fonte di dati di (pre)formazione

Sopra, abbiamo immaginato che i professionisti utilizzino la classifica dei benchmark come guida per la selezione dei modelli, ma poi addestrino i modelli selezionati da zero sulle proprie fonti di dati (spesso proprietarie). Ma i professionisti spesso possono andare oltre e lo fanno.

In alcuni casi, potrebbe essere possibile eseguire l'addestramento su un set di dati di riferimento e utilizzare direttamente il modello risultante nella propria applicazione. Ciò dipende dal dominio e dall'attività ed è più adatto quando lo spostamento della distribuzione è minimo e l'insieme delle etichette delle classi è stabile. Ad esempio, è ragionevole distribuire un riconoscitore di cifre preaddestrato su MNIST, ma non altrettanto un classificatore di immagini preaddestrato su ILSVRC (senza qualche tipo di adattamento al dominio di destinazione). In effetti, ILSVRC è costituito da un sottoinsieme piuttosto arbitrario di 1.000 classi di ImageNet e un modello preaddestrato è corrispondentemente limitato nell'insieme di etichette che è in grado di produrre. L'ImmagineNet

¹La difficoltà di accertare fino a che punto i risultati di uno studio si generalizzano oltre la popolazione studiata tormenta tutte le scienze statistiche. Vedi, ad esempio.⁵¹³

Il progetto Roulette è stato una dimostrazione significativa di ciò che accade quando un modello addestrato sul set di dati (completo) di ImageNet viene applicato a una diversa distribuzione di test, costituita principalmente da immagini di persone. I risultati furono grotteschi. La dimostrazione è stata interrotta, ma molti risultati archiviati possono essere trovati in articoli sul progetto.⁵¹⁴ Infine, consideriamo un set di dati di riferimento del sistema di raccomandazioni. Non c'è modo nemmeno di tentare di utilizzarli direttamente come dati di addestramento perché è altamente improbabile che gli utenti su cui si desidera fare previsioni siano presenti nel set di addestramento.

Nella maggior parte dei casi, i creatori di set di dati di benchmark non intendono utilizzarli come fonte di dati di addestramento, sebbene i set di dati di benchmark vengano spesso utilizzati in modo improprio per questo scopo. Una rara eccezione è The Pile: un ampio corpus di testi in inglese (800 GB) esplicitamente mirato alla formazione di modelli linguistici. Per migliorare le capacità di generalizzazione dei modelli addestrati su questo corpus, gli autori hanno incluso testi diversi provenienti da 22 fonti diverse.⁵¹⁵

Anche quando i set di dati di riferimento non sono utili come dati di formazione per i motivi sopra menzionati, possono essere utili come dati pre-formazione per il trasferimento dell'apprendimento. Il trasferimento dell'apprendimento si riferisce all'utilizzo di un modello esistente come punto di partenza per la costruzione di un nuovo modello. Potrebbe essere necessario un nuovo modello perché la distribuzione dei dati è cambiata rispetto a ciò per cui era stato ottimizzato il modello esistente o perché mira a risolvere un compito completamente diverso. Ad esempio, un modello pre-addestrato su ImageNet (o ILSVRC) può essere adattato tramite ulteriore formazione per riconoscere specie diverse (spostamento di distribuzione) o come parte di un modello di didascalia di immagini (un compito diverso).

Esistono diverse intuizioni per spiegare perché il trasferimento dell'apprendimento è spesso efficace. Il primo è che gli strati finali di una rete neurale corrispondono a rappresentazioni semanticamente di alto livello dell'input. La pre-formazione è un modo per apprendere queste rappresentazioni che tendono ad essere utili per molti compiti. Un'altra intuizione è che il pre-addestramento è un modo di inizializzare i pesi che offre un miglioramento rispetto all'inizializzazione casuale in quanto richiede meno campioni dal dominio target per la convergenza.

La formazione preliminare offre il vantaggio pratico di poter condividere la conoscenza contenuta in un set di dati senza rilasciare i dati grezzi. Molti set di dati, soprattutto quelli creati dalle aziende che utilizzano i dati dei clienti, non possono essere pubblicati per motivi di privacy o riservatezza. Il rilascio di modelli preaddestrati è quindi un'importante via di condivisione delle conoscenze dall'industria al mondo accademico. La condivisione di modelli preaddestrati è utile anche per gli utenti per i quali la formazione da zero ha costi proibitivi. Tuttavia, le preoccupazioni relative alla privacy e alla protezione dei dati emergono nel contesto della condivisione di modelli preaddestrati a causa della possibilità che i dati personali utilizzati per la formazione possano essere ricostruiti dal modello preaddestrato.⁵¹⁶ Concludiamo la nostra analisi dei ruoli dei set di dati di riferimento.

Abbiamo identificato sei ruoli distinti: (1) fornire dati campionati da distribuzioni che si verificano nel mondo reale che consentono indagini in gran parte indipendenti dal dominio degli algoritmi di apprendimento; (2) consentire progressi specifici del dominio fornendo set di dati rappresentativi delle attività del mondo reale in quel dominio, eliminando tuttavia i dettagli non necessari; (3) fornire un metodo numerico conveniente, anche se rozzo, per monitorare il progresso scientifico su un problema; (4) consentire il confronto dei modelli e consentire ai professionisti di esternalizzare la selezione dei modelli alle classifiche pubbliche; (5) fornire una fonte di dati pre-formazione per

apprendimento della rappresentazione, inizializzazione del peso, ecc.; (6) fornire una fonte di dati di formazione. La progressione di questi sei ruoli è generalmente verso una maggiore specificità del dominio e del compito e dall'orientamento scientifico all'orientamento pratico.

Le basi scientifiche dei benchmark di machine learning

Ora esaminiamo un apparente mistero: se e perché l'approccio benchmark funziona nonostante la pratica di test ripetuti sugli stessi dati.

Metodologicamente, gran parte della pratica moderna del machine learning si basa su una variante di tentativi ed errori, che chiamiamo paradigma del train-test. I professionisti costruiscono ripetutamente modelli utilizzando un numero qualsiasi di euristiche e testano le loro prestazioni per vedere cosa funziona. Per quanto riguarda l'addestramento, tutto va bene, soggetto solo a vincoli computazionali, purché le prestazioni sembrino buone durante i test. Prova ed errore sono validi fintanto che il protocollo di test è sufficientemente robusto da assorbire la pressione su di esso. Esamineremo in che misura ciò è vero nel machine learning.

Da un punto di vista teorico, il modo migliore per testare le prestazioni di un classificatore è raccogliere un nuovo set di dati sufficientemente ampio e calcolare l' errore medio su quel set di test. La raccolta dei dati, tuttavia, è un compito difficile e costoso. Nella maggior parte delle applicazioni, i professionisti non possono campionare nuovi dati per ciascun modello che vorrebbero provare. Una pratica diversa è quindi diventata lo standard de facto.

I professionisti dividono solitamente il proprio set di dati in due parti, un set di addestramento utilizzato per addestrare un modello e un set di test utilizzato per valutarne le prestazioni.² Spesso la suddivisione viene determinata al momento della creazione del set di dati. I set di dati utilizzati per i benchmark in particolare hanno una suddivisione fissa persistente nel tempo. Un certo numero di variazioni su questo tema vanno sotto il nome di metodo holdout.

Le competizioni di machine learning hanno adottato lo stesso formato. L'azienda Kaggle, ad esempio, dalla sua fondazione ha organizzato centinaia di concorsi.

In una competizione, un set di resistenze viene tenuto segreto e viene utilizzato per classificare i partecipanti in una classifica pubblica durante lo svolgimento della competizione. Alla fine, il vincitore finale è colui che ottiene il punteggio più alto in un set di test segreto separato non utilizzato fino a quel momento.

In tutte le applicazioni del metodo di controllo la speranza è che il set di test serva come nuovo campione che fornisca buone stime delle prestazioni per tutti i modelli. Il problema centrale è che i professionisti non utilizzano i dati dei test solo una volta per poi ritirarli immediatamente dopo. I dati di test vengono utilizzati in modo incrementale per costruire un modello alla volta incorporando il feedback ricevuto in precedenza dai dati di test. Ciò porta al timore che alla fine i modelli inizino ad adattarsi eccessivamente ai dati di test.

Questo tipo di overfitting è talvolta chiamato overfitting adattivo o overfitting human-in-the-loop.

Duda, Hart e Stork riassumono adeguatamente il problema nel loro libro di testo del 1973: 517

Nei primi lavori sul riconoscimento di modelli, quando gli esperimenti venivano spesso condotti con un numero molto piccolo di campioni, gli stessi dati venivano spesso utilizzati per progettare e testare il classificatore. Questo errore è frequente

²A volte i professionisti dividono i propri dati in più suddivisioni, ad esempio formazione, convalida e test insieme. Tuttavia, per la nostra discussione qui, ciò non sarà necessario.

denominato “test sui dati di addestramento”. Un problema correlato ma meno ovvio sorge quando un classificatore viene sottoposto a una lunga serie di perfezionamenti guidati dai risultati di test ripetuti sugli stessi dati. Questa forma di “formazione sui dati di prova” spesso sfugge all’attenzione finché non vengono ottenuti nuovi campioni di prova.

Quasi mezzo secolo dopo, Hastie, Tibshirani e Friedman mettono ancora in guardia nell’edizione del 2017 del loro influente libro di testo:⁵¹⁸

Idealmente, il set di test dovrebbe essere conservato in un “caveau” ed essere tirato fuori solo alla fine dell’analisi dei dati. Supponiamo invece di utilizzare ripetutamente il test-set, scegliendo il modello con l’errore più piccolo. Quindi l’errore del test set del modello finale scelto sottostimerà il vero errore del test, a volte in modo sostanziale.

Sebbene il suggerimento di conservare i dati dei test in un “archivio” sia sicuro, non potrebbe essere più lontano dalla realtà della pratica moderna. I set di dati di test più diffusi spesso vedono decine di migliaia di valutazioni.

Eppure non sembra che si stia verificando un overfitting adattivo. Ricordiamo i grafici a dispersione di Recht et al. sopra: i grafici ammettono un adattamento lineare pulito con pendenza positiva. In altre parole, quanto migliore è un modello sul vecchio set di test, tanto migliore sarà sul nuovo set di test.

Ma si noti che i modelli più recenti, cioè quelli con prestazioni più elevate sul set di test originale, hanno avuto più tempo per adattarsi al set di test e per incorporare più informazioni al riguardo. Tuttavia, quanto migliore è il modello eseguito sul vecchio set di prova, tanto migliore sarà il suo rendimento sul nuovo set. Inoltre, su CIFAR-10 vediamo chiaramente che il calo assoluto delle prestazioni diminuisce con l’aumentare della precisione sul vecchio set di test. In particolare, se il nostro obiettivo era quello di fare bene sul nuovo set di test, apparentemente la nostra migliore strategia è continuare a progredire sul vecchio set di test.

La comprensione teorica del motivo per cui la pratica dell’apprendimento automatico non ha portato a un overfitting sta ancora recuperando terreno. Qui evidenziamo una delle tante possibili spiegazioni, chiamata principio della classifica. È un effetto sottile in cui i pregiudizi di pubblicazione costringono i ricercatori a perseguire risultati all’avanguardia e a pubblicare modelli solo se vedono miglioramenti significativi rispetto ai modelli precedenti. Questa pratica culturale può essere formalizzata dall’algoritmo Ladder di Blum & Hardt. Per ciascun classificatore, confronta l’errore di controllo del classificatore con l’errore di controllo precedentemente più piccolo ottenuto da qualsiasi classificatore incontrato finora. Se l’errore è inferiore al migliore precedente di un certo margine, annuncia l’errore di controllo del classificatore corrente e lo rileva come il migliore visto finora. È importante sottolineare che se l’errore non è inferiore di un margine, l’algoritmo rilascia il migliore precedente (anziché il nuovo errore). Si può dimostrare che l’algoritmo Ladder evita il sovraccarico, nel senso che misura accuratamente l’errore del classificatore più performante tra quelli incontrati.⁵¹⁹

Prassi e cultura di riferimento

La discussione di cui sopra suggerisce l’importanza delle pratiche culturali per una piena comprensione dei set di dati di riferimento. Parliamo ora di questi in modo più dettagliato,

evidenziando sia i creatori dei set di dati che gli utenti. Queste pratiche hanno contribuito al successo dell'approccio orientato ai benchmark, ma hanno anche avuto un impatto sui danni associati ai dati. Cominciamo con i creatori.

I creatori di benchmark definiscono il compito. Ciò comporta, tra le altre cose, la selezione del problema di alto livello, la definizione della variabile target, la procedura per campionare i dati e la funzione di punteggio. Se è necessaria l'annotazione manuale dei dati, il creatore del set di dati deve sviluppare un codice o una rubrica per farlo e orchestrare il lavoro collettivo, se necessario. Di solito è necessaria la pulizia dei dati per garantire etichette di alta qualità .

Nel definire il compito, gli sviluppatori di benchmark si trovano a dover affrontare un difficile equilibrio: un compito considerato troppo facile utilizzando le tecniche esistenti non stimolerà l'innovazione, mentre un compito considerato troppo difficile potrebbe essere demotivante. Trovare il punto giusto richiede esperienza, giudizio e un po' di fortuna. Se viene raggiunto il giusto equilibrio, il benchmark guida i progressi sul problema. In questo modo, i creatori di benchmark svolgono un ruolo enorme nel definire la visione e l'agenda per le comunità di machine learning. È noto che la selezione dei compiti nei benchmark influisce sulla classificazione dei modelli, il che influenza e influenza la direzione del progresso nella comunità.⁵²⁰ Questo effetto potrebbe diventare più pronunciato nel tempo a causa della crescente concentrazione su un numero inferiore di set di dati.⁵²¹ Come esempio di il tipo di decisioni che gli sviluppatori di benchmark devono prendere e il modo in cui influenzano la direzione della ricerca, considera MNIST. Come discusso in precedenza, è stato derivato da un precedente set di dati rilasciato dal NIST

in cui il set di training e test proveniva da fonti diverse, ma il MNIST ha eliminato questo spostamento di distribuzione. I creatori del MNIST hanno sostenuto che ciò era necessario perché

Trarre conclusioni sensate dagli esperimenti di apprendimento richiede che il risultato sia indipendente dalla scelta del set di addestramento e dal test tra il set completo di campioni.

In altre parole, se un algoritmo funziona bene sul NIST, non è chiaro quanto ciò sia dovuto alla sua capacità di apprendere la distribuzione di addestramento e quanto sia dovuto alla sua capacità di ignorare le differenze tra le distribuzioni del treno e quelle dei test.

MNIST consente ai ricercatori di concentrarsi selettivamente sulla prima domanda. Questo approccio è stato fruttuoso nel 1995. Decenni dopo, quando problemi come la classificazione MNIST sono stati effettivamente risolti, l'attenzione dei creatori di set di dati di riferimento si è rivolta ai metodi per gestire il cambiamento della distribuzione che LeCun et al. ha scelto giustamente di ignorarlo.⁵²² Un altro equilibrio difficile è tra

l'astrarre i dettagli del dominio in modo che il compito sia accessibile a un'ampia fascia di esperti di machine learning, e il preservare abbastanza dettagli in modo che i metodi che funzionano nell'impostazione del benchmark si traducano in impostazioni di produzione. Uno dei motivi per cui il Premio Netflix è stato così popolare è perché i dati sono solo una matrice ed è possibile ottenere buone prestazioni (nel senso di superare il livello di base di Netflix) senza pensare veramente al significato dei dati. Non era necessaria alcuna comprensione della psicologia del cinema o dell'utente, né utile, come si è scoperto. È possibile che la competenza nel settore si sarebbe rivelata essenziale se

il problema era stato formulato in modo diverso, ad esempio per richiedere una spiegazione o per enfatizzare le buone prestazioni anche per gli utenti con pochissime valutazioni precedenti.

Un'altra sfida per i creatori di set di dati è evitare perdite. In una storia apocrifa degli albori della visione artificiale, un classificatore veniva addestrato a discriminare tra immagini di carri armati russi e americani con una precisione apparentemente elevata, ma si scoprì che ciò era dovuto solo al fatto che i carri armati russi erano stati fotografati in una giornata nuvolosa e quelli americani in una giornata soleggiata.⁵²³ La perdita di dati si riferisce a una relazione spuria tra il vettore delle caratteristiche e la variabile target che è un artefatto della raccolta dati o della strategia di campionamento. Poiché la relazione spuria non sarà presente quando il modello viene distribuito, la perdita di solito porta a stime gonfiate delle prestazioni del modello. Kaufman et al. presentare una panoramica delle perdite nell'apprendimento automatico.⁵²⁴ Un'altra responsabilità

fondamentale dei creatori di set di dati di riferimento è quella di implementare un quadro di prova del treno. La maggior parte dei concorsi prevede varie restrizioni nel tentativo di impedire sia il sovraadattamento accidentale al set di test della classifica sia il reverse engineering intenzionale. Anche se, come abbiamo descritto sopra, la prassi di benchmark differisce dalla versione del libro di testo del metodo di controllo, i professionisti sono arrivati a una serie di tecniche che hanno funzionato nella pratica, anche se la nostra comprensione teorica del perché funzionano è ancora in fase di recupero.

Facendo un passo indietro, in qualsiasi attività scientifica ci sono i difficili compiti di inquadrare il problema, garantire che i metodi abbiano validità interna ed esterna e interpretare i risultati. I creatori di set di dati di benchmark gestiscono il maggior numero possibile di questi compiti difficili, semplificando l'obiettivo degli utenti del set di dati al punto che, se un ricercatore supera le prestazioni dello stato dell'arte, c'è una buona probabilità che ci sia una visione scientifica da qualche parte. i metodi, anche se l'estrazione di questa intuizione potrebbe ancora richiedere lavoro. A semplificare ulteriormente le cose per gli utenti del set di dati è il fatto che non ci sono restrizioni oltre ai vincoli computazionali su come il ricercatore utilizza i dati di addestramento, purché le prestazioni sul set di test siano buone.

Per essere chiari, questo approccio presenta molte insidie. I ricercatori raramente eseguono i test di ipotesi statistica necessari per avere fiducia nell'affermazione che un modello funziona meglio di un altro.⁵²⁵ La nostra comprensione di come tenere conto delle numerose fonti di varianza in queste misurazioni delle prestazioni è ancora in evoluzione; un documento del 2021 che mira a farlo sostiene che molte delle affermazioni sulle prestazioni all'avanguardia nelle prestazioni del linguaggio naturale e nella visione artificiale non reggono quando sottoposte a tali test.⁵²⁶ Da tempo esistono articoli

che sottolineano le limitazioni di ciò che ricercatori e professionisti possono imparare dalla valutazione delle prestazioni benchmark.^{527,528} David Aha, co-creatore dell'archivio UCI, ricorda che queste limitazioni erano ben comprese già nel 1995, solo pochi anni dopo la creazione dell'archivio.⁵²⁹

Sebbene sia importante riconoscere i limiti, vale anche la pena sottolineare che questo approccio funziona. Uno dei motivi di questo successo è che le questioni scientifiche riguardano principalmente gli algoritmi e non le popolazioni da cui vengono campionati i set di dati.

In effetti, è opportuno sostenere che altre comunità scientifiche dovrebbero adottare l'approccio della comunità dell'apprendimento automatico, a volte chiamato Common

Metodo del compito.481 Diversi campi scientifici, tra cui l'economia, le scienze politiche, la psicologia, la genetica e molti altri, hanno visto un'infusione di metodi di apprendimento automatico insieme a una nuova attenzione alla massimizzazione dell'accuratezza predittiva come obiettivo di ricerca. Questi cambiamenti sono stati accompagnati da un'ondata di fallimenti in termini di riproducibilità, con grandi frazioni di articoli pubblicati che sono caduti preda di trappole come la fuga di dati.530 L'utilizzo dell'approccio basato sui dataset di riferimento avrebbe potuto evitare la maggior parte di queste trappole.

Passiamo ora agli utenti del set di dati. Gli utenti del benchmark hanno abbracciato la libertà offerta da questo approccio. Di conseguenza, la comunità di utenti è ampia: ad esempio, la piattaforma di data science Kaggle conta oltre 5 milioni di utenti registrati, di cui oltre 130.000 hanno partecipato a un concorso. C'è meno controllo nella ricerca sull'apprendimento automatico che in altre discipline. Molti risultati importanti eludono la peer review. Se una tecnica si comporta bene in classifica, si ritiene che parli da sola. Molte persone che contribuiscono a questi risultati non sono formalmente affiliate agli istituti di ricerca.

Nel complesso, la cultura del progresso nell'apprendimento automatico combina la cultura dell'erudizione accademica, dell'ingegneria e persino del gioco, con una comunità di hobbisti e professionisti che condividono suggerimenti e trucchi sui forum e si impegnano in competizioni amichevoli. Questa cultura a ruota libera può sembrare sconcertante per alcuni osservatori, soprattutto data la sensibilità di alcuni dei set di dati coinvolti. La mancanza di controlli significa minori opportunità di formazione etica.

C'è un altro aspetto della cultura del benchmark che amplifica i danni associati ai dati: raccogliere dati senza consenso informato e distribuirli ampiamente senza un contesto adeguato. Molti set di dati moderni, soprattutto nel campo della visione artificiale e dell'elaborazione del linguaggio naturale, vengono recuperati dal web. In questi casi non è possibile ottenere il consenso informato dei singoli autori dei contenuti.

Che ne dici di un set di dati come il Premio Netflix in cui un'azienda rilascia dati dalla propria piattaforma? Anche se le aziende rivelano nei loro termini di servizio che i dati potrebbero essere utilizzati per la ricerca, è dubbio che sia stato ottenuto il consenso informato poiché pochi utenti leggono e comprendono i documenti dei Termini di servizio e a causa della complessità delle questioni coinvolte.

Quando i dati di un individuo diventano parte di un set di dati di riferimento, vengono ampiamente distribuiti. I set di dati di benchmark più diffusi vengono scaricati da migliaia di ricercatori, studenti, sviluppatori e hobbisti. Le norme scientifiche richiedono inoltre che i dati siano conservati a tempo indeterminato nell'interesse della trasparenza e della riproducibilità. Pertanto, non solo i singoli dati in questi set di dati potrebbero essere distribuiti e visualizzati ampiamente, ma vengono visualizzati in una forma che li priva del loro contesto originale. Una battuta di cattivo gusto scritta sui social e poi cancellata potrebbe essere catturata insieme ai documenti della Biblioteca del Congresso.

Danni associati ai dati

Ora discuteremo alcuni importanti tipi di dati associati ai set di dati di riferimento e come mitigarli. Non intendiamo implicare che tutti questi danni

sono “colpa” dei creatori di set di dati, ma comprendere il ruolo dei dati in questi danni porterà chiarezza su come intervenire.

Danni a valle e rappresentativi

I danni a valle di un set di dati sono quelli che derivano dai modelli addestrati su di esso. Questo è un tipo di danno che viene subito in mente: dati errati possono portare a modelli errati che possono causare danni alle persone che presumibilmente servono. Ad esempio, i sistemi distorti di previsione del rischio criminale danneggiano in modo sproporzionato, tra gli altri, le popolazioni nere, appartenenti alle minoranze e a quelle sottoposte a un eccesso di polizia.

Le proprietà dei set di dati che a volte (ma non sempre, e non in modi facilmente prevedibili) si propagano a valle includono squilibrio, pregiudizi, stereotipi e categorizzazione. Per squilibrio intendiamo una rappresentanza ineguale dei diversi gruppi.

Ad esempio, Buolamwini e Gebru hanno sottolineato che due benchmark di analisi facciale, IJB-A e Adience, presentavano in stragrande maggioranza soggetti dalla pelle più chiara.³³⁹ Per bias del set di dati intendiamo associazioni errate, in particolare quelle corrispondenti a pregiudizi sociali e storici. Ad esempio, un set di dati che misura gli arresti come indicatore della criminalità può riflettere i pregiudizi della polizia e le leggi discriminatorie. Per stereotipi intendiamo associazioni che riflettono accuratamente una proprietà del mondo (o di una specifica cultura in un determinato momento) che si ritiene essere il risultato di pregiudizi sociali e storici. Ad esempio, le associazioni di genere e di occupazione possono essere chiamate stereotipi. Per categorizzazione intendiamo l’assegnazione di etichette discrete (spesso binarie) ad aspetti complessi dell’identità come il genere e la razza.

I danni rappresentazionali si verificano quando i sistemi rafforzano la subordinazione di alcuni gruppi lungo le linee dell’identità. I danni rappresentazionali potrebbero essere danni a valle – come quando i modelli applicano etichette offensive a persone appartenenti ad alcuni gruppi – ma potrebbero essere inerenti al set di dati. Ad esempio, ImageNet contiene numerosi insulti ed etichette offensive ereditate da WordNet e immagini pornografiche di persone che non hanno acconsentito alla loro inclusione nel set di dati.^{531, 532} Mentre i danni a valle e rappresentativi sono due categorie che hanno attirato molta attenzione e

critica, ci sono molti altri danni che spesso insorgono, tra cui il costo ambientale dei modelli di addestramento su set di dati inutilmente grandi⁵³³ e la cancellazione del lavoro dei soggetti che hanno contribuito ai dati⁵²⁹ o degli annotatori che li hanno etichettati.⁴⁹⁵ Per una panoramica delle preoccupazioni etiche associate ai set di dati, vedere l’indagine di Paullada et al.⁵³⁴

Mitigare i danni: una panoramica

Gli approcci per mitigare i danni associati ai dati si stanno rapidamente sviluppando. Qui esaminiamo alcune idee selezionate.

Un approccio prende di mira il fatto che molti set di dati di machine learning sono scarsamente documentati e spesso mancano i dettagli sulla loro creazione. Ciò porta a una serie di problemi che vanno dalla mancanza di riproducibilità e preoccupazioni sulla validità scientifica all’uso improprio e alle preoccupazioni etiche. In risposta, i fogli dati per i set di dati sono un modello

e iniziativa di Gebru et al. promuovere annotazioni più dettagliate e sistematiche per i set di dati.⁵³⁵ Una scheda dati richiede che il creatore di un set di dati risponda a domande relative a diverse aree di interesse: motivazione, composizione, processo di raccolta, preelaborazione/pulizia/etichettatura, usi, distribuzione, manutenzione. Uno degli obiettivi è che il processo di creazione di una scheda dati aiuti ad anticipare le questioni etiche relative al set di dati. Ma le schede tecniche mirano anche a rendere le pratiche relative ai dati più riproducibili e ad aiutare i professionisti a selezionare fonti di dati più adeguate.

Andando oltre i datasheet, Jo e Gebru⁵³⁶ traggono lezioni dalle scienze archivistiche e bibliotecarie per la costruzione e la documentazione di set di dati di machine learning. Queste lezioni attirano l'attenzione sulle questioni relative al consenso, all'inclusività, al potere, alla trasparenza, all'etica e alla privacy.

Altri approcci rimangono nel paradigma della raccolta dati minimamente curata, ma mirano a modificare o ripulire i contenuti ritenuti problematici nei set di dati. I creatori di ImageNet si sono impegnati a rimuovere insulti e termini dannosi, nonché categorie considerate non immaginabili o che non possono essere caratterizzate utilizzando immagini.

“Vegetariano” e “filantropo” sono due di queste categorie che sono state rimosse.⁵³² Lo strumento REVISE mira ad automatizzare parzialmente il processo di identificazione di vari tipi di bias nei set di dati visivi.⁵³⁷

Mitigare i danni separando i ruoli dei set di dati

La nostra analisi dei diversi ruoli svolti dai set di dati consente una maggiore chiarezza nel mitigare i danni preservando i benefici. Questa analisi non è intesa come alternativa ai numerosi approcci già proposti per mitigare i danni. Piuttosto, può affinare il nostro pensiero e rafforzare altre strategie di mitigazione del danno.

La nostra osservazione principale è che il riutilizzo di set di dati di benchmark scientifici nelle pipeline di ingegneria complica gli sforzi per affrontare pregiudizi e danni. I tentativi di affrontare i danni derivanti da tali set di dati a duplice uso lasciano i creatori con un enigma. Da un lato, i set di dati di benchmark devono essere di lunga durata: molti set di dati di benchmark creati decenni fa continuano ad essere utili e ampiamente utilizzati oggi. Pertanto, la modifica di un set di dati in futuro quando vengono conosciuti nuovi danni ne comprometterà l'utilità scientifica, poiché le prestazioni del set di dati modificato potrebbero non essere significativamente paragonabili alle prestazioni del set di dati più vecchio.

D'altro canto, tentare di anticipare tutti i possibili danni durante la creazione del set di dati non è fattibile se il set di dati verrà utilizzato come dati di addestramento o pre-addestramento. L'esperienza dimostra che i set di dati si rivelano utili per una serie di attività a valle in continua espansione, alcune delle quali non erano nemmeno state concepite al momento della creazione del set di dati.

Sono possibili compromessi migliori se esiste una chiara separazione tra parametri di riferimento scientifici e set di dati orientati alla produzione. Nei casi in cui lo stesso set di dati può essere potenzialmente utile per entrambi gli scopi, i creatori dovrebbero prendere in considerazione la creazione di due versioni o fork dei dati, perché molte delle strategie di mitigazione del danno che si applicano a una non si applicano all'altra e viceversa.

Per imporre questa separazione, i creatori di set di dati di benchmark dovrebbero considerare di evitare l'uso del set di dati nelle pipeline di produzione vietandolo esplicitamente nei termini

di utilizzo. Attualmente le licenze di molti set di dati di riferimento vietano gli usi commerciali. Questa restrizione ha un effetto simile, ma non è il modo migliore per fare questa distinzione. Dopotutto, i modelli di produzione possono non essere commerciali: possono essere costruiti da ricercatori o governi, e quest'ultima categoria ha un potenziale di danno particolarmente elevato. Allo stesso tempo, vietare gli usi commerciali è probabilmente troppo severo, in quanto vieta l'uso del set di dati come guida per la selezione del modello, un uso che non solleva gli stessi rischi di danni a valle.

Uno dei motivi per cui esistono interventi di equità applicabili ai dataset di benchmark scientifici ma non ai dataset di produzione è che, come abbiamo sostenuto, la maggior parte dell'utilità scientifica dei benchmark è catturata dalle prestazioni relative dei modelli.

Il fatto che gli interventi che danneggiano la performance assoluta possano essere accettabili offre un maggiore margine di manovra per gli sforzi di mitigazione del danno. Considerare i parametri di classificazione delle immagini . Ipotizziamo che la classificazione relativa dei modelli sarà influenzata solo in minima parte se il set di dati viene modificato per rimuovere tutte le immagini contenenti persone (mantenendo le stesse proprietà di alto livello incluso il numero di classi e immagini). Un intervento di questo tipo eviterebbe un'ampia gamma di danni associati ai set di dati, preservandone al contempo gran parte dell'utilità scientifica.

Al contrario, uno dei motivi per cui esistono interventi di equità applicabili ai set di dati di produzione ma non parametri di riferimento scientifici è che gli interventi per i set di dati di produzione possono essere fortemente guidati dalla comprensione dei loro impatti a valle in applicazioni specifiche. Il linguaggio e le immagini, in particolare, catturano una tale varietà di stereotipi culturali che eliminarli tutti si è rivelato impossibile.⁵³⁸ È molto più semplice progettare interventi una volta fissata un'applicazione e il contesto culturale in cui verrà implementata. Allo stesso set di dati utilizzato in applicazioni diverse possono essere applicabili interventi diversi . A differenza dei benchmark scientifici, la standardizzazione dei set di dati non è necessaria in ambito ingegneristico.

In effetti, il miglior luogo di intervento anche per le distorsioni dei dati potrebbe essere a valle dei dati. Ad esempio, è stato osservato per molti anni che i sistemi di traduzione online perpetuano gli stereotipi di genere quando si traducono pronomi di genere neutro. Il testo "O bir dottore. O bir hem,sire." può essere tradotto dal turco all'inglese come "È un dottore. Lei è un'infermiera." Google Translate ha mitigato questo problema mostrando più traduzioni in questi casi.^{539, 540} Rispetto agli interventi sui dati, questo ha il vantaggio di rendere il potenziale bias (o, in alcuni casi, la traduzione errata) più visibile all'utente.

La nostra analisi indica molte aree in cui ulteriori ricerche potrebbero aiutare a chiarire le implicazioni etiche. In particolare, il ruolo di pre-addestramento dei set di dati di benchmark occupa un'area grigia in cui non è chiaro quando e in che misura i bias dei dati si propagano all'attività/dominio target. La ricerca in questo settore è agli inizi;⁵⁴¹ questa ricerca è vitale perché l'uso (abusivo) di parametri di riferimento scientifici per la pre-formazione nelle linee di produzione è oggi comune ed è improbabile che cessi nel prossimo futuro.

I set di dati non dovrebbero essere visti come artefatti tecnici statici e neutrali. I dati che potrebbero derivare da un set di dati dipendono non solo dal suo contenuto ma anche dalle regole, dalle norme e dalla cultura che ne circondano l'utilizzo. Pertanto, la modifica di queste pratiche culturali è un potenziale modo per mitigare i danni. Come discusso in precedenza, la mancanza di conoscenza del dominio da parte degli utenti del set di dati è diventata quasi una virtù della macchina

apprendimento. Questo atteggiamento dovrebbe essere riconsiderato poiché tende ad accettare i punti ciechi etici.

I set di dati richiedono la gestione, da parte del creatore del set di dati o di un'altra entità o insieme di entità designate. Consideriamo il problema dei derivati: i set di dati di benchmark più diffusi sono spesso estesi da altri ricercatori con caratteristiche aggiuntive, e questi set di dati derivati possono introdurre la possibilità di danni non presenti nell'originale (nella stessa misura). Ad esempio, il set di dati dei volti Labeled Faces in the Wild (LFW) è stato annotato da altri ricercatori con caratteristiche come razza, genere e attrattiva.^{542, 543} Indipendentemente dall'etica della LFW stessa, il set di dati derivato consente nuove applicazioni che classificano le persone dall'apparenza in modi dannosi.³ Naturalmente, non tutti i derivati sono eticamente problematici. Giudicare e far rispettare tali distinzioni etiche è possibile solo se esiste un meccanismo di governance in atto.

Oltre i set di dati

In questa sezione finale, discutiamo importanti questioni scientifiche ed etiche che sono rilevanti per i set di dati ma vanno anche oltre i set di dati, pervadendo l'apprendimento automatico: validità, inquadramento dei problemi e limiti alla previsione.

Lezioni dalla misurazione

La teoria della misurazione è una scienza consolidata con radici antiche. In breve, la misurazione consiste nell'assegnare numeri agli oggetti nel mondo reale in modo da riflettere le relazioni tra questi oggetti. La misurazione traccia un'importante distinzione tra un costrutto che desideriamo misurare e la procedura di misurazione che abbiamo utilizzato per creare una rappresentazione numerica del costrutto.

Ad esempio, possiamo pensare a un esame di matematica ben progettato come a una misurazione delle capacità matematiche di uno studente. Ci si aspetta che uno studente con maggiori capacità matematiche di un altro ottenga un punteggio più alto nell'esame. Visto in questo modo, un esame è una procedura di misurazione che assegna numeri agli studenti. L'abilità matematica di uno studente è il costrutto che speriamo di misurare. Desideriamo che l'ordinamento di questi numeri rifletta l'ordinamento degli studenti in base alle loro abilità matematiche. Una procedura di misurazione rende operativo un costrutto.

Ogni problema di previsione ha una variabile obiettivo, la cosa che stiamo cercando di prevedere.⁴ Considerando la variabile obiettivo come un costrutto, possiamo applicare la teoria della misurazione per capire cosa rende una variabile obiettivo buona.

La scelta di una variabile target inadeguata non può essere risolta con dati aggiuntivi. Infatti, maggiore è la quantità di dati che inseriamo nel nostro modello, migliore sarà la sua capacità di catturare la variabile target difettosa. Nemmeno una migliore qualità o diversità dei dati rappresenta una cura.

³Lo scopo previsto del set di dati derivato è consentire la ricerca di corpora di immagini di volti per attributi descrivibili.

⁴Ricordiamo che in un problema di previsione abbiamo covariate X da cui stiamo cercando di prevedere a variabile Y. Questa variabile Y è ciò che chiamiamo variabile obiettivo nel nostro problema di previsione.

Tutti i criteri formali di equità che coinvolgono la variabile obiettivo, di cui separazione e sufficienza sono due esempi importanti⁵, sono privi di significato o addirittura fuorvianti quando la variabile obiettivo stessa è il luogo della discriminazione.

Ma cosa rende una variabile target buona o cattiva? Cerchiamo di capire meglio questo questione considerando alcuni esempi.

1. Prevedere il valore dell'indice Standard and Poor 500 (S&P 500) alla chiusura della Borsa di New York domani.
2. Prevedere se un individuo andrà in default su un prestito.
3. Prevedere se un individuo commetterà un crimine.

Il primo esempio è piuttosto innocuo. Fa riferimento a una variabile target abbastanza solida, anche se si basa su una serie di fatti sociali.

Il secondo esempio è un'applicazione comune della modellistica statistica che è alla base di gran parte del moderno credit scoring negli Stati Uniti. A prima vista un evento predefinito sembra una variabile target netta. Ma la realtà è diversa. In un dataset pubblico della Federal Reserve¹¹⁶ gli eventi di default sono codificati da una cosiddetta variabile di performance che misura una grave inadempienza in almeno una linea di credito in un determinato periodo di tempo. Più specificamente, il rapporto della Federal Reserve afferma che

la misura si basa sulla performance di conti nuovi o esistenti e misura se gli individui sono in ritardo di 90 giorni o più su uno o più dei loro conti o se avevano un elemento di registro pubblico o un nuovo conto presso l'agenzia di riscossione durante il periodo di prestazione.⁶

Il nostro terzo esempio incontra il problema di misurazione più preoccupante. Come determiniamo se un individuo ha commesso un crimine? Ciò che possiamo determinare con certezza è se un individuo è stato arrestato o meno e giudicato colpevole di un crimine.

Ma ciò dipende in modo cruciale da chi sarà probabilmente il primo a essere sorvegliato e da chi sarà in grado di manovrare con successo il sistema di giustizia penale dopo un arresto.

Stabilire quale sia una buona variabile target, in tutta la sua generalità, può coinvolgere l'intero apparato della teoria della misurazione. Lo scopo della teoria della misurazione, tuttavia, va oltre la definizione di variabili target affidabili e valide per la previsione. La misurazione entra in gioco ogni volta che creiamo funzionalità per un problema di apprendimento automatico e dovrebbe quindi essere una parte essenziale del processo di creazione dei dati.⁵⁴⁴

Giudicare la qualità di una procedura di misurazione è un compito difficile. La teoria della misurazione ha due importanti quadri concettuali per discutere su ciò che rende buona la misurazione. Uno è l'affidabilità. L'altro è la validità.

L'affidabilità descrive le differenze osservate in più misurazioni dello stesso oggetto in condizioni identiche. Considerando la variabile di misurazione come una variabile casuale, l'affidabilità riguarda la varianza tra indipendenti in modo identico

⁵Ricordiamo dal Capitolo 3 che la separazione richiede che l'attributo protetto sia indipendente dalla previsione condizionata sulla variabile obiettivo. La sufficienza richiede che la variabile obiettivo sia indipendente dall'attributo protetto data la previsione.

⁶Citazione dal rapporto della Federal Reserve.

misurazioni distribuite. In quanto tale, l'affidabilità può essere paragonata alla nozione statistica di varianza.

La validità riguarda quanto bene la procedura di misurazione, in linea di principio, cattura il concetto che cerchiamo di misurare. Se l'affidabilità è analoga alla varianza, si è tentati di vedere la validità come analoga alla distorsione. Ma la situazione è un po' più complicata. Non esiste un criterio formale semplice che potremmo utilizzare per stabilire la validità. In pratica, la validità si basa in larga misura sulla competenza umana e sui giudizi soggettivi.

Un approccio per formalizzare la validità è chiedersi quanto bene un punteggio predice alcuni criteri esterni. Questa si chiama validità esterna. Ad esempio, potremmo giudicare una misura dell'affidabilità creditizia in base alla sua capacità di prevedere il default in uno scenario di prestito. Sebbene la validità esterna porti a criteri tecnici concreti, identifica essenzialmente una buona misurazione con accuratezza predittiva. Tuttavia, la validità non si limita certamente a questo.

La validità di costrutto è un quadro per discutere la validità che include numerosi diversi tipi di prove. Messick evidenzia sei aspetti della validità di costrutto:

- Contenuto: quanto bene è il contenuto dello strumento di misura, ad esempio gli elementi di un questionario, misurano il costrutto di interesse? • Sostanziale: il costrutto è supportato da solide basi teoriche? • Strutturale: il punteggio esprime relazioni nel dominio dei costrutti? • Generalizzabilità: il punteggio è generalizzabile tra popolazioni diverse, impostazioni e attività?
- Esterno: il punteggio prevede con successo i criteri esterni? • Consequenziale: quali sono i rischi potenziali derivanti dall'utilizzo del punteggio in merito al pregiudizio, all'equità e alla giustizia distributiva?

Di questi diversi criteri, la validità esterna è quello più familiare ai professionisti del machine learning. Ma la pratica del machine learning farebbe bene ad abbracciare anche gli altri criteri, più qualitativi. Il criterio consequenziale è stato controverso, ma Messick difende con forza la sua inclusione come aspetto di validità.⁵⁴⁵ In definitiva, la misurazione ci costringe a confrontarci con una domanda spesso sorprendentemente scomoda: cosa stiamo cercando di fare quando prevediamo qualcosa?

Inquadramento del problema: confronti con l'uomo

Un'ambizione di lunga data della ricerca sull'intelligenza artificiale è quella di eguagliare o superare le capacità cognitive umane mediante un algoritmo. Questo desiderio porta spesso a confronti tra esseri umani e macchine su vari compiti. I giudizi sull'accuratezza umana spesso entrano anche nel dibattito su quando utilizzare modelli statistici in contesti decisionali ad alta posta in gioco.

Il confronto tra decisori umani e modelli statistici non è affatto nuovo. Per decenni i ricercatori hanno confrontato l'accuratezza dei giudizi umani con quella dei modelli statistici.⁵⁴⁶

Anche nell'ambito dell'apprendimento automatico, il dibattito risale a molto tempo fa. Un articolo del 1991 di Bromley e Sackinger ha confrontato esplicitamente le prestazioni delle reti neurali artificiali con una misura della precisione umana sul set di dati delle cifre USPS che precede i famosi dati MNIST.⁴⁹¹ Un primo esperimento ha messo la precisione umana al 2,5%, un secondo esperimento ha trovato la numero 1,51%, mentre un terzo ha riportato il numero 2,37%.⁵⁴⁷ Il confronto con le cosiddette linee di base

umane è diventato da allora ampiamente accettato nella comunità del machine learning. La Electronic Frontier Foundation (EFF), ad esempio, ospita un importante archivio di misure di progresso dell'IA che confronta le prestazioni dei modelli di apprendimento automatico con la precisione umana riportata su numerosi benchmark.

Per i dati ILSVRC 2012, la precisione umana riportata è del 5,1%.⁷ Questo numero spesso citato corrisponde alle prestazioni di un singolo annotatore umano che è stato "addestrato su 500 immagini e ha annotato 1500 immagini di prova".⁴⁹⁴ Un secondo annotatore che è stato "addestrato su 100 immagini e poi annotate 258 immagini di prova" ha raggiunto una precisione del 12%. Sulla base di questo numero del 5,1%, i ricercatori hanno annunciato nel 2015 che il loro modello era "il primo a superare le prestazioni a livello umano".⁵⁴⁸ Non sorprende che questa affermazione abbia ricevuto un'attenzione significativa da parte dei media.

Tuttavia, una successiva indagine più attenta sulla "precisione umana" su ImageNet ha rivelato un quadro molto diverso.⁵⁴⁹ I ricercatori hanno scoperto che solo i modelli del 2020 sono effettivamente alla pari con il più potente etichettatore umano. Inoltre, limitando i dati a 590 classi di oggetti su 1000 classi in totale, il miglior etichettatore umano ha ottenuto risultati molto migliori con un errore inferiore all'1% anche rispetto ai migliori modelli predittivi. Ricordiamo che i dati ILSVRC del 2012 comprendevano solo 118 diverse razze di cani, alcune delle quali sono estremamente difficili da distinguere per chiunque non sia un esperto di cani addestrato. In effetti, i ricercatori hanno dovuto consultarsi con esperti dell'American Kennel Club (AKC) per chiarire i casi difficili di diverse razze canine.

La semplice rimozione delle sole classi di cani aumenta le prestazioni del miglior etichettatore umano con un errore inferiore all'1,3%.

C'è un altro fatto preoccupante. Piccole variazioni nel protocollo di raccolta dati risultano avere un effetto significativo sulle prestazioni dei classificatori automatici: "i punteggi di precisione anche dei migliori classificatori di immagini sono ancora altamente sensibili alle minuzie del processo di pulizia

dei dati".⁵¹⁰ Questi risultati non mettono in dubbio solo su come misuro l'accuratezza umana, ma anche sulla validità del presunto costrutto teorico di "accuratezza umana" stessa. Tuttavia, la comunità del machine learning ha adottato un approccio piuttosto casuale per misurare l'accuratezza umana. Molti ricercatori presumono che il costrutto dell'accuratezza umana esista in modo inequivocabile e sia qualunque numero emerga da un protocollo di test ad hoc per un gruppo di esseri umani. Questi protocolli ad hoc spesso danno luogo a confronti aneddotici di discutibile valore scientifico.

Giudizi non validi sulle prestazioni umane rispetto alle macchine non sono solo un errore scientifico, ma hanno anche il potenziale di creare narrazioni che supportano scelte politiche sbagliate in questioni politiche ad alta posta in gioco sull'uso di strumenti predittivi.

⁷Per essere precisi, questo numero si riferisce alla frazione di volte in cui l'etichetta dell'immagine corretta non era contenuta nelle prime 5 etichette previste del modello o dell'essere umano.

modelli nelle decisioni consequenziali. Ad esempio, la politica della giustizia penale è guidata dall'affermazione secondo cui i metodi statistici sono superiori ai giudici nel prevedere il rischio di recidiva o di mancata comparizione in tribunale. Tuttavia, questi confronti sono dubbi perché i giudici non risolvono puri problemi di previsione ma piuttosto incorporano altri fattori come l'indulgenza nei confronti degli imputati più giovani.²

Inquadramento del problema: concentrarsi su un unico obiettivo di ottimizzazione

I problemi della vita reale raramente implicano l'ottimizzazione di un singolo obiettivo e più comunemente comportano una sorta di compromesso tra più obiettivi. Il modo migliore per formulare questo problema come problema di ottimizzazione statistica è sia un'arte che una scienza. Tuttavia, le attività di benchmark, in particolare quelle con classifiche, tendono a scegliere un unico obiettivo. Per i benchmark di alto profilo, il conseguente "adattamento eccessivo alla formulazione del problema" può provocare punti ciechi scientifici e limitare l'applicabilità dei risultati pubblicati a contesti pratici.

Ad esempio, all'epoca in cui Netflix lanciò il Premio, era ben noto che la raccomandazione non è solo una questione di massimizzazione dell'accuratezza predittiva e, anche nella misura in cui lo è, non esiste un'unica misura che sia sempre appropriata.⁵⁵⁰ Eppure il concorso focalizzato esclusivamente sull'accuratezza della previsione valutata da una singola metrica. Alcuni anni dopo la fine del concorso, Netflix ha rivelato che la maggior parte del lavoro inserito nella classifica non si era tradotto in modelli di produzione. In parte il motivo era che il concorso non riusciva a cogliere la gamma degli obiettivi e dei vincoli di Netflix: la stretta dipendenza dei consigli dall'interfaccia utente; il fatto che gli "utenti" sono tipicamente nuclei familiari composti da membri con gusti diversi; spiegabilità; freschezza e molto altro ancora.⁵⁵¹ Se molti dei dati ricavati dalla classifica non si possono nemmeno generalizzare al contesto produttivo di Netflix,

il divario tra Netflix e le altre piattaforme orientate alla raccomandazione è molto maggiore. In particolare, come piattaforma cinematografica, Netflix è insolita in quanto ha un inventario relativamente statico rispetto a quelli con contenuti generati dagli utenti come YouTube o Facebook. Quando il pool di contenuti è dinamico, è necessaria una classe diversa di algoritmi. L'attrazione esercitata dal Premio Netflix sulla ricerca sui sistemi di raccomandazione potrebbe aver distolto l'attenzione da quest'ultimo tipo di algoritmo per molti anni, anche se è difficile saperlo con certezza perché il controfattuale non è osservabile.

Le gare formali di machine learning, anche se causano punti ciechi dovuti alla necessità di scegliere un unico obiettivo di ottimizzazione, sono almeno attentamente strutturate per promuovere il progresso scientifico in un senso stretto. Probabilmente più dannose sono le competizioni informali che sembrano inevitabilmente emergere in presenza di un importante set di dati di riferimento, con conseguenti esiti sfortunati, come il rifiuto di articoli interessanti perché non sono riusciti a superare lo stato dell'arte, o la pubblicazione di articoli non originali perché lo hanno fatto. battere lo stato dell'arte attraverso l'applicazione (scientificamente insignificante) di una maggiore potenza di calcolo.

Un altro svantaggio di un campo orientato alla ricerca unidimensionale e competitiva è che diventa strutturalmente difficile affrontare i pregiudizi nei modelli e nei classificatori. Se un concorrente adotta misure per impedire la propagazione di errori nel set di dati

ai loro modelli, ci sarà un calo di accuratezza (perché l'accuratezza viene giudicata su un set di dati distorto) e meno persone presteranno attenzione al lavoro.

Man mano che le questioni relative all'equità nell'apprendimento automatico hanno guadagnato importanza, i set di dati di benchmark incentrati sull'equità sono proliferati, come il Pilot Parliamentarians Benchmark per l'analisi facciale³³⁹ e l'Equity Evaluation Corpus per l'analisi del sentimento.³³⁶ Un vantaggio di questo approccio è che i dati scientifici e culturali Il meccanismo dell'innovazione orientata ai benchmark può essere riproposto per la ricerca sull'equità. Un potenziale pericolo è la legge di Goodhart, che afferma, nella sua forma più ampia, "Quando una misura diventa un obiettivo, cessa di essere una buona misura". Come abbiamo sottolineato in questo libro, l'equità è multiforme e i benchmark possono catturare solo nozioni ristrette di equità. Sebbene questi possano essere utili strumenti diagnostici, se vengono interpretati erroneamente come obiettivi a sé stanti, la ricerca focalizzata sull'ottimizzazione di questi parametri di riferimento potrebbe non comportare equità in un senso più sostanziale. Inoltre, la costruzione di questi set di dati è stata spesso casuale, senza un'adeguata attenzione alle questioni di validità.⁵⁵² Oltre a creare

benchmark incentrati sull'equità, la comunità dell'equità algoritmica ha anche riproposto i benchmark precedenti verso lo studio delle questioni di equità. Consideriamo il set di dati del censimento dal repository UCI discusso in precedenza.

Originariamente ha guadagnato popolarità come fonte di dati del mondo reale. Il suo utilizzo è accettabile per studiare questioni algoritmiche come, ad esempio, la forza relativa degli alberi decisionali e la regressione logistica. Ci aspettiamo che le risposte siano insensibili a questioni come il contesto culturale dei dati. Ma ora viene utilizzato per studiare questioni di equità, ad esempio il modo in cui l'accuratezza della classificazione tende a variare in base alla razza o al genere. **Per tali domande, le risposte sono sensibili ai dettagli delle sottopopolazioni.** Inoltre, il compito di classificazione associato al benchmark (previsione del reddito trattato come una variabile binaria) è artificiale e non corrisponde ad alcuna applicazione nella vita reale. Pertanto, le disparità di accuratezza (e altre misurazioni legate all'equità) potrebbero apparire diverse per un compito diverso, o se i dati fossero stati campionati in modo diverso, o se provenissero da un tempo o da un luogo diverso. L'utilizzo di set di dati di riferimento per fare affermazioni generalizzabili sull'equità richiede un'attenzione particolare alle questioni di contesto, campionamento e validità. Bao et al. chiedersi se i dati di riferimento per i sistemi socio-tecnici come la giustizia penale siano utili. Sottolineano che la cultura del benchmark – in cui l'attenzione è sui metodi, con il set di dati secondario e il contesto ignorato – è in contrasto con le effettive esigenze di equità e giustizia, dove l'attenzione al contesto è fondamentale.⁵⁵³

Limiti dei dati e previsione

L'apprendimento automatico fallisce in molti scenari ed è importante comprendere i casi di fallimento tanto quanto le storie di successo.

La Fragile Families Challenge era una competizione di machine learning basata sullo studio Fragile Families and Child Wellbeing (FFCWS).⁵⁵⁴ Partendo da un campione casuale di nascite in ospedale tra il 1998 e il 2000, la FFCWS ha seguito migliaia di famiglie americane nel corso di 15 anni, raccogliere informazioni dettagliate sui figli delle famiglie, sui loro genitori, sui risultati scolastici e

l'ambiente sociale più ampio. Una volta che la famiglia ha accettato di partecipare allo studio, i dati sono stati raccolti alla nascita del bambino e poi all'età di 1, 3, 5, 9 e 15 anni.

La Fragile Families Challenge si è conclusa nel 2017. Il set di dati alla base del concorso contiene 4242 righe, una per ogni famiglia, e 12943 colonne, una per ciascuna variabile più un numero ID di ciascuna famiglia. Delle 12942 variabili, 2358 sono costanti (cioè hanno lo stesso valore per tutte le righe), principalmente a causa di oscuramenti per questioni di privacy ed etica. Dei circa 55 milioni (4242×12942) voci del set di dati, circa il 73% non ha valore. I valori mancanti hanno molte possibili ragioni, tra cui la mancata risposta delle famiglie intervistate, l'abbandono dei partecipanti allo studio, nonché relazioni logiche tra caratteristiche che implicano la mancanza di alcuni campi a seconda di come sono impostati gli altri. Ci sono sei variabili di risultato, misurate all'età di 15 anni: 1) media dei voti del bambino (GPA), 2) determinazione del bambino, 3) sfratto della famiglia, 4) difficoltà materiali della famiglia , 5) licenziamento del caregiver e 6) partecipazione del caregiver alla formazione professionale. .

L'obiettivo della competizione era prevedere il valore delle variabili di risultato all'età di 15 anni sulla base dei dati da 1 a 9 anni. Come è comune per le competizioni, la sfida prevedeva una suddivisione dei dati in tre direzioni: allenamento, classifica e set di test. Il set di formazione è pubblicamente disponibile per tutti i partecipanti, i dati della classifica supportano una classifica durante tutta la competizione e il set di test viene utilizzato per determinare un vincitore finale.

L'esito della sfida di previsione è stato deludente. Anche i modelli vincenti hanno ottenuto risultati appena migliori rispetto a una semplice linea di base, le loro previsioni non differivano molto rispetto alla previsione della media di ciascun risultato.

Cosa ha causato le scarse prestazioni del machine learning sui dati delle famiglie fragili? Una possibilità ovvia è che nessuno dei concorrenti abbia trovato le tecniche di apprendimento automatico giuste per questo compito. Ma il fatto che 160 team di esperti motivati abbiano presentato migliaia di modelli nel corso di cinque mesi rende ciò altamente improbabile. Inoltre, modelli provenienti da classi di modelli disparate hanno fatto tutte previsioni molto simili (e ugualmente errate), suggerendo che gli algoritmi di apprendimento non rappresentavano il limite.⁸ Esistono alcune altre possibilità tecniche che potrebbero spiegare le prestazioni deludenti, tra cui la dimensione del campione, lo studio progettazione e i valori mancanti.

Ma c'è anche una ragione più fondamentale che rimane plausibile. Forse le dinamiche delle traiettorie di vita sono intrinsecamente imprevedibili nel corso del ritardo di sei anni tra la misurazione delle covariate e la misurazione del risultato.

Questo divario di sei anni, ad esempio, includeva la Grande Recessione, un periodo di shock economici e declino tra il 2007 e il 2009, che avrebbe potuto cambiare le traiettorie in modi imprevedibili.

In effetti, c'è una ragione importante per cui anche le prestazioni dei modelli nella sfida, per quanto sconfortanti, potrebbero sopravvalutare ciò che possiamo aspettarci in un contesto reale. Questo perché ai modelli è stato permesso di sbirciare nel futuro, per così dire. I set di training e test sono stati estratti dalla stessa distribuzione e, in particolare, dallo stesso periodo di tempo, come è la pratica standard nell'apprendimento automatico

⁸Ciò evidenzia un vantaggio dell'approccio del set di dati benchmark rispetto a uno con meno standardizzazione : anche quando non si riescono a fare progressi sostanziali sulla previsione, possiamo imparare qualcosa di prezioso da quel fallimento.

ricerca. Pertanto, i dati incorporano già informazioni sugli effetti della Grande Recessione e di altri shock globali durante questo periodo. In un'applicazione reale, i modelli devono essere addestrati su dati del passato mentre le previsioni riguardano il futuro. Pertanto, c'è sempre qualche deriva: un cambiamento nella relazione tra le covariate e il risultato. Ciò pone un ulteriore limite alle prestazioni del modello.

L'apprendimento automatico funziona meglio in un mondo statico e stabile in cui il passato assomiglia al futuro. La sola previsione può essere una scelta sbagliata quando anticipiamo cambiamenti dinamici o quando cerchiamo di ragionare sull'effetto che azioni ipotetiche avrebbero nel mondo reale.

Riepilogo

I set di dati di benchmark sono fondamentali per l'apprendimento automatico. Svolgono molti ruoli, tra cui consentire l'innovazione algoritmica, misurare i progressi e fornire dati di formazione. Dalla sua sistematizzazione alla fine degli anni '80, la valutazione delle prestazioni sui benchmark è gradualmente diventata una pratica onnipresente perché rende più difficile per i ricercatori imbrogliare intenzionalmente o meno.

Ma un'eccessiva attenzione ai benchmark comporta molti inconvenienti. I ricercatori dedicano sforzi prodigiosi all'ottimizzazione dei modelli per ottenere prestazioni all'avanguardia. I risultati sono spesso poco interessanti dal punto di vista scientifico e di scarsa rilevanza per i professionisti perché i parametri di riferimento omettono molti dettagli del mondo reale. L'approccio amplifica inoltre i danni associati ai dati, inclusi danni a valle, danni rappresentativi e violazioni della privacy.

Mentre scriviamo questo libro, l'approccio benchmark viene messo sotto esame a causa di queste preoccupazioni etiche. Sebbene i vantaggi e gli svantaggi dei benchmark siano entrambi ben noti, il nostro obiettivo generale in questo capitolo è stato quello di fornire un unico quadro di riferimento che possa aiutare ad analizzarli entrambi. La nostra posizione è che vale la pena preservare il nucleo dell'approccio basato sui benchmark, ma immaginiamo un futuro in cui i benchmark svolgeranno un ruolo più modesto come uno dei tanti modi per far avanzare la conoscenza. Per mitigare i danni associati ai dati, riteniamo che siano necessari cambiamenti sostanziali alle pratiche di creazione, utilizzo e governance dei set di dati. Abbiamo delineato alcuni modi per farlo, ampliando la letteratura emergente su questo argomento.

Note del capitolo

Questo capitolo è stato sviluppato e pubblicato per la prima volta da Hardt e Recht nel libro di testo *Patterns, Predictions, and Actions: Foundations of Machine Learning*.¹¹⁹ Con il permesso degli autori, riportiamo qui gran parte del testo originale con solo lievi modifiche. Abbiamo rimosso una quantità significativa di materiale sull'analisi adattiva dei dati e sul problema dell'adattamento eccessivo nei benchmark di apprendimento automatico. Abbiamo aggiunto nuovo materiale sul ruolo svolto dai set di dati, nonché discussioni sull'equità e sulle preoccupazioni etiche relative ai set di dati.

L'adattività nel riutilizzo degli holdout è stata studiata da Dwork et al.⁵⁵⁵ e sono stati condotti lavori successivi nell'area dell'analisi adattiva dei dati. Preoccupazioni simili svaniscono

il nome dell'inferenza dopo la selezione nella comunità statistica.

La raccolta e l'uso di grandi insiemi di dati ad hoc (un tempo definiti "big data") sono stati esaminati in diversi lavori importanti, si veda, ad esempio, Boyd e Crawford,⁵⁵⁶ così come Tufekci.⁵⁵⁷,⁵⁵⁸ Più recentemente, Couldry e Mejias⁵⁵⁹ utilizzano il termine colonialismo dei dati per enfatizzare i processi attraverso i quali i dati vengono appropriati e le comunità emarginate vengono sfruttate attraverso la raccolta dei dati. Olteanu et al.⁵⁶⁰ discutono pregiudizi, trappole metodologiche e questioni etiche nel contesto dell'analisi dei dati sociali. In particolare, l'articolo fornisce tassonomie di pregiudizi e problemi che possono sorgere nell'approvvigionamento, raccolta, elaborazione e analisi dei dati sociali. Il testo classico di Bowker e Star spiega perché la categorizzazione è un'attività moralmente carica.¹⁹¹ Per una discussione sui danni dei sistemi di categorie incorporati nei set di dati di apprendimento automatico, vedere *Atlas of AI*.⁵⁶¹ I vantaggi

dell'approccio del set di dati benchmark sono discussi in un discorso di Mark Liberman, che lo definisce il metodo delle attività comuni.⁵⁶² Paullada, Raji, Ben-der, Denton e Hanna analizzano lo sviluppo dei set di dati e i casi d'uso nella ricerca sull'apprendimento automatico.⁵⁶³ Un'indagine condotta da Fabris, Messina, Silvello e Susto elenca e discute numerosi usi di set di dati nella letteratura sull'equità.⁵⁶⁴ Denton, Hanna, Amironesei, Smart e Nicole forniscono una genealogia di ImageNet attraverso una lente critica.⁵⁶⁵ Raji, Bender, Paullada, Denton e Hanna forniscono una panoramica delle preoccupazioni derivanti dal basare la nostra comprensione del progresso su una piccola raccolta di parametri di riferimento influenti.⁵⁶⁶ Il progetto EFF AI metrics è disponibile all'indirizzo: <https://www.eff.org/ai/metrics>.

Per un'introduzione alla teoria della misurazione, non specifica per le scienze sociali, vedere i libri di Hand.⁵⁶⁷,⁵⁶⁸ Il libro di testo di Bandalos⁵⁶⁹ si concentra sulle applicazioni alle scienze sociali, compreso un capitolo sull'equità. Liao, Taori, Raji e Schmidt forniscono una tassonomia degli errori di valutazione in molti sottocampi dell'apprendimento automatico , comprendendo questioni di validità sia interna che esterna.⁵⁷⁰

Bibliografia

- ¹ Gates, Susan Wharton, Perry, Vanessa Gail e Zorn, Peter M. 2002. "Sottoscrizione automatizzata nei prestiti ipotecari: buone notizie per i meno serviti?" Dibattito sulla politica abitativa, 13(2):369–391.
- ² Stevenson, Megan T e Doleac, Jennifer L. 2022. "Valutazione algoritmica del rischio nelle mani degli esseri umani". Disponibile presso SSRN.
- 3 Ingold, David e Soper, Spencer. 2016. "Amazon non considera la razza dei suoi clienti. dovrebbe?" <https://www.bloomberg.com/graphics/2016-amazon-same-day/>.
- 4 Crawford, Kate. 2013. "I pregiudizi nascosti nei big data". Harvard Business Review, 1.
- 5 Kaggle. 2012. "The Hewlett Foundation: punteggio automatizzato dei saggi". <https://www.kaggle.com/c/al-più-presto.6>
- Hanna, Rema N e Linden, Leigh L. 2012. "Discriminazione nella classificazione". American Economic Journal: Politica economica, 4(4):146–68.
- 7 Spietsma, Maresa. 2013. "Discriminazioni nell'assegnazione dei voti: evidenze sperimentali da parte degli insegnanti della scuola primaria". Economia empirica, 45(1):523–538.
- 8 Ashkenas, Jeremy, Park, Haeyoun e Pearce, Adam. 2017. "Anche con l'azione affermativa, i neri e gli ispanici sono più sottorappresentati nelle università più prestigiose rispetto a 35 anni fa". <https://www.nytimes.com/interactive/2017/08/24/us/affirmative-action.html> .
- 9 Barcas, Solon e Selbst, Andrew D.. 2016. "L'impatto disparato dei big data". Revisione della legge della California, 104.
- ¹⁰ Plaugic, Lizzie. 2017. "Il creatore di Faceapp si scusa per il filtro 'caldo' schiarente della pelle dell'app". Il limite. <https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology>.
- ¹¹ Manthorpe, Rowland. 2017. "Il concorso di bellezza dei robot beauty.ai è tornato". Cablato nel Regno Unito. <https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai> .
- ¹² Corey, Ethan e Lo, Puck. 2019. "L'errore della 'mancata apparizione'". L'appello.
- 13 Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. e Fei-Fei, L.. 2009. "ImageNet: un database di immagini gerarchiche su larga scala" . Nell'Proc. CVPR.
- 14 Miller, George A. 1995. "Wordnet: un database lessicale per l'inglese". Comunicazioni dell'ACM, 38(11):39–41.
- 15 Birhane, Abeba e Prabhu, Vinay Uday. 2021. "Set di dati di immagini di grandi dimensioni: una vittoria di Pirro per la visione artificiale?" Nel 2021 Conferenza invernale IEEE sulle applicazioni della visione artificiale (WACV), pagine 1536–1546. IEEE.
- ¹⁶ Roth, Lorna. 2009. "Guardando Shirley, la norma ultima: equilibrio del colore, tecnologie dell'immagine ed equità cognitiva". Giornale canadese di comunicazione, 34(1).
- 17 Torralba, Antonio e Efros, Alexei A. 2011. "Unbiased look at dataset bias". Nell'Proc. CVPR, pagine 1521–1528. IEEE.
- ¹⁸ Liu, Zicheng, Zhang, Cha e Zhang, Zhengyou. 2007. "Miglioramento della qualità dell'immagine percettiva basato sull'apprendimento per le videoconferenze". In Multimedia and Expo, Conferenza internazionale IEEE 2007 , pagine 1035–1038. IEEE.
- 19 Kaufman, Liad, Lischinski, Dani, e Werman, Michael. 2012. "Miglioramento automatico delle foto in base al contenuto ". In Computer Graphics Forum, volume 31, pagine 2528–2540. Biblioteca in linea Wiley.
- ²⁰ Kalantari, Nima Khademi e Ramamoorthi, Ravi. 2017. "Imaging ad alta gamma dinamica profonda di scene dinamiche". ACM Trans. Grafico, 36(4):144.
- ²¹ Bonham, Vence L, Callier, Shawneequa L e Royal, Charmaine D. 2016. "La medicina di precisione ci porterà oltre la razza?" Il giornale di medicina del New England, 374(21):2003.

- ²² Wilson, James F, Weale, Michael E, Smith, Alice C, Gratrix, Fiona, Fletcher, Benjamin, Thomas, Mark G, Bradman, Neil e Goldstein, David B. 2001. "Struttura genetica della popolazione della risposta variabile ai farmaci". *Genetica della natura*, 29(3):265.
- 23 Caliskan, Aylin, Bryson, Joanna J. e Narayanan, Arvind. 2017. "La semantica deriva automaticamente dai corpora linguistici contengono pregiudizi di tipo umano". *Scienza*, 356(6334):183–186.
- 24 Danesi, Marcel. 2014. *Dizionario dei media e delle comunicazioni*. Routledge.
- 25 Hardt, Moritz. 2014. "Quanto sono ingiusti i big data". **quanto-** <https://medium.com/@mrtz/i-big-data-sono-ingiusti-9aa544d739de>.
- ²⁶ Caruana, Rich, Lou, Yin, Gehrke, Johannes, Koch, Paul, Sturm, Marc e Elhadad, Noemie. 2015. "Modelli intelligibili per l'assistenza sanitaria: previsione del rischio di polmonite e riammissione ospedaliera a 30 giorni". Nell'Proc. 21° ACM SIGKDD, pagine 1721–1730.
- 27 Joachims, Thorsten, Swaminathan, Adith e Schnabel, Tobias. 2017. "Apprendimento imparziale per classificare con feedback distorto". Nell'Proc. 10a conferenza internazionale sulla ricerca sul Web e sul data mining, pagine 781–789. ACM.
- ²⁸ Sweeney, Latanya. 2013. "Discriminazione nell'erogazione di annunci online". *Coda*, 11(3):10:10–10:29.
- 29 Dobbie, Will, Goldin, Jacob e Yang, Crystal. 2016. "Gli effetti della custodia cautelare sulla condanna, sulla criminalità futura e sull'occupazione: prove da giudici assegnati in modo casuale". *Rapporto tecnico*, Ufficio nazionale di ricerca economica.
- 30 Lamù, Kristian e Isaac, William. 2016. "Prevedere e servire?" *Significato*, 13(5):14–19.
- 31 Ensign, Danielle, Friedler, Sorelle A, Neville, Scott, Scheidegger, Carlos e Venkatasubramanian, Suresh. 2017. "Circuiti di feedback incontrollati nella polizia predittiva". arXiv prestampa [arXiv:1706.09847](https://arxiv.org/abs/1706.09847).
- 32 Zhang, Junzhe e Bareinboim, Elias. 2018. "Equità nel processo decisionale: la formula di spiegazione causale". Nell'Proc. 32° AAAI.
- 33 Roccia, David e Grant, Heidi. 2016. "Perché i team diversificati sono più intelligenti". *Revisione aziendale di Harvard*. <https://hbr.org/2016/11/why-diverse-teams-are-smarter>.
- 34 Freeman, Richard B e Huang, Wei. 2015. "Collaborare con persone come me: coautore etnico-spedire negli Stati Uniti". *Giornale di economia del lavoro*, 33 (S1): S289–S318.
- 35 Baroca, Solone. 2014. "Mettere i dati al lavoro". In *Virginia Eubanks Seeta Peña Gangadharan e Solon Barcas (a cura di), Data and Discrimination: Collected Essays*, pagine 59–62. Fondazione Nuova America .
- 36 Jackson, John W. e VanderWeele, Tyler J.. 2018. "Analisi di decomposizione per identificare obiettivi di intervento per ridurre le disparità". *Epidemiologia*, pagine 825–835.
- 37 Kay, Matthew, Matuszek, Cynthia e Munson, Sean A. 2015. "Rappresentazione ineguale e stereotipi di genere nei risultati di ricerca di immagini per le occupazioni". Nell'Proc. 33a conferenza sui fattori umani nei sistemi informatici, pagine 3819–3828. ACM.
- 38 Crawford, Kate. 2017. "Il problema dei pregiudizi". NeurIPS Keynote https://www.youtube.com/watch?v=fMym_BKWQzk .
- 39 Huszár, Ferenc, Ktena, Sofia Ira, O'Brien, Conor, Belli, Luca, Schlaikjer, Andrew, and Hardt, Moritz. 2022. "Amplificazione algoritmica della politica su Twitter". *Atti dell'Accademia Nazionale delle Scienze*, 119(1):e2025334119.
- 40 Reisman, Dillon, Schultz, Jason, Crawford, Kate e Whittaker, Meredith. 2018. "Valutazioni d'impatto algoritmiche: un quadro pratico per la responsabilità delle agenzie pubbliche". <https://ainowinstitute.org/aiareport2018.pdf> .
- 41 Munoz, Cecilia, Smith, Megan e Patil, D. 2016. "Big data: un rapporto sui sistemi algoritmici, opportunità e diritti civili". Ufficio esecutivo del Presidente. La Casa Bianca.
- 42 Campolo, Alex, Sanfilippo, Madelyn, Whittaker, Meredith e Crawford, Kate. 2017. "Ai adesso Rapporto 2017". AI Now Institute della New York University.
- 43 Friedman, Batya e Nissenbaum, Helen. 1996. "Bias nei sistemi informatici". *Transazioni ACM attive Sistemi informativi (TOIS)*, 14(3):330–347.
- 44 Pedreschi, Dino, Ruggieri, Salvatore, and Turini, Franco. 2008. "Dati consapevoli della discriminazione minerario". Nell'Proc. 14° SIGKDD. ACM.
- 45 Pasquale, Frank. 2015. *La società della scatola nera: gli algoritmi segreti che controllano denaro e informazioni*. Stampa dell'Università di Harvard.
- 46 O'Neil, Cathy. 2016. *Armi di distruzione della matematica: come i big data aumentano la disuguaglianza e minacciano la democrazia*. Libri di Broadway.

- 47 Eubanks, Virginia. 2018. Automatizzare la disuguaglianza: come gli strumenti high-tech profilano, sorvegliano e puniscono i poveri. Stampa di San Martino.
- 48 Nobile, Safiya Umoja. 2018. Algoritmi di oppressione: come i motori di ricerca rafforzano il razzismo. New York Premere.
- 49 Konger, Kate e Chen, Brian X.. 2022. "Un cambiamento da parte di Apple sta tormentando le società Internet, in particolare Meta". Il New York Times <https://www.nytimes.com/2022/02/03/technology/apple-privacy-changes-meta.html>.
- 50 Richardson , William Jamal. 2017. "Contro l'inclusione dei neri nel riconoscimento facciale".
- 51 Powles, Julia e Nissenbaum, Helen. 2018. "Il seducente diversivo del 'risolvere' pregiudizi nell'intelligenza artificiale ". Medio.
- 52 Weber, Max. 2019. Economia e società. Harvard University Press, Cambridge, MA.
- 53 Strandburg, Katherine J.. 2019. "Regolamentazione e strumenti decisionali automatizzati imperscrutabili". Columbia Law Review, 119(7):1851–1886.
- 54 Creel, Kathleen e Hellman, Deborah. 2021. "Il levitano algoritmico: arbitrarietà, equità e opportunità nei sistemi decisionali algoritmici". Documento di ricerca sul diritto pubblico e sulla teoria giuridica della Virginia , (2021-2013).
- 55 Kroll, Joshua A., Huey, Joanna, Baracas, Solon, Felten, Edward W., Reidenberg, Joel R., Robinson, David G. e Yu, Harlan. 2017. "Algoritmi responsabili". Revisione giuridica dell'Università della Pennsylvania, 165(3):633–705.
- 56 Cedro, Danielle Keats. 2008. "Il giusto processo tecnologico". Revisione giuridica dell'Università di Washington, 85(6):1249–1313.
- 57 Christie, James. 2020. "L'orizzonte postale è lo scandalo e la presunzione di attendibilità delle prove informatiche". Prove digitali ed elettricità Firma L. Rev., 17:49.
- 58 Kaplow, Louis. 1992. "Regole contro standard: un'analisi economica". Duca Giurisprudenza Journal, 42(3):557–629.
- 59 Alkhatib, Ali e Bernstein, Michael. 2019. "Algoritmi a livello di strada: una teoria sul divario tra politica e decisioni". In Conferenza sui fattori umani nei sistemi informatici (CHI), pagine 1–13. 2018. cura".
- 60 Colin Lecher. salute "Cosa succede quando un algoritmo taglia <https://www.theverge.com/2018/3/21/17144260/> il tuo algoritmo-medico-sanitario-arkansas-paralisi-cerebrale.
- ⁶¹ Clarke, Roger. 1988. "Informatica e dataveglianza". Comunicazioni dell'ACM, 31(5):498–512.
- ⁶² Kaminski, Margot E. e Urban, Jennifer M.. 2021. "Il diritto di contestare l'intelligenza artificiale". Columbia Law Review, 121(7):1957–2048.
- 63 Gilman, Michele. 2020. "Algoritmi legislativi sulla povertà". Rapporto tecnico, Data & Society, New York, NY.
- 64 Nissenbaum, Helen. 1996. "La responsabilità in una società informatizzata". Scienza e ingegneria Etica, 2(1):25–42.
- 65 Binns, Reuben, Kleek, Max Van, Veale, Michael, Lyngs, Ulrik, Zhao, Jun e Shadbolt, Nigel. 2018. "È ridurre un essere umano a una percentuale": percezioni di giustizia nelle decisioni algoritmiche". Atti della conferenza CHI 2018 sui fattori umani nei sistemi informatici, pagine 1–14.
- ⁶⁶ Collins, Harry. 1991. Esperti artificiali: conoscenza sociale e macchine intelligenti. MIT Press, Cambridge, MA.
- 67 Forsythe, Diana E.. 2000. Studiare coloro che ci studiano: un antropologo nel mondo dell'artificiale Intelligenza. Stanford University Press, Stanford, CA.
- ⁶⁸ Hand, David J.. 2006. "La tecnologia dei classificatori e l'illusione del progresso". Scienze statistiche, 21(1):1–14.
- 69 Burrell, Jenna. 2016. "Come la macchina 'pensa': comprendere l'opacità nell'apprendimento automatico algoritmi". Big Data e società, 3(1).
- 70 Ribeiro, Marco Tulio, Singh, Sameer, e Guestrin, Carlos. 2016. ""Perché dovrei fidarmi di te?": Spiegare le previsioni di qualsiasi classificatore". Atti della 22a conferenza internazionale ACM SIGKDD sulla scoperta della conoscenza e sul data mining, pagine 1135–1144.
- 71 Perelman, Les. 2012. "Costruire validità, lunghezza, punteggio e tempo nelle valutazioni di scrittura classificate olisticamente: il caso contro il punteggio automatizzato dei saggi (aes)". Progressi internazionali nella ricerca scritta: culture, luoghi, misure, pagine 121–131.

- 72 Johnson, Rebecca Ann e Zhang, Simone. 2022. "Che cos'è il controfattuale burocratico? Priorità categorica rispetto a quella algoritmica nella politica sociale statunitense". pagine 1671–1682.
- 73 Abebe, Rediet, Barcas, Solon, Kleinberg, Jon, Levy, Karen, Raghavan, Manish e Robinson, David G. 2020. "Roles for computing in social change". pagine 252–260.
- 74 Tyler, Tom R. 1988. "Che cos'è la giustizia procedurale?: criteri utilizzati dai cittadini per valutare l'equità delle procedure legali". *Law & Society Review*, 22(1):103–135.
- 75 Passi, Samir e Barcas, Solone. 2019. "Formulazione dei problemi ed equità". Nella Conferenza del Equità, responsabilità e trasparenza, pagine 39–48.
- 76 Hand, David J.. 1994. "Deconstructing Statistical Questions". Giornale della Royal Statistical Society: Serie A (Statistiche nella società), 157(3):317–338.
- 77 Richardson, Rashida. 2022. "Segregazione razziale e società basata sui dati: come la nostra incapacità di fare i conti con le cause profonde perpetua realtà separate e disuguali". *Berkeley Technology Law Journal*, 36(3):1051–1090.
- 78 Lamù, Kristian e Isaac, William. 2016. "Prevedere e servire?" *Significato*, 13(5):14–19.
- 79 Harcourt, Bernard E. 2008. Contro la previsione: profilazione, polizia e punizione in un'era attuariale. Stampa dell'Università di Chicago.
- ⁸⁰ Gandy, Oscar H.. 2010. "Coinvolgere la discriminazione razionale: esplorare le ragioni per imporre vincoli normativi sui sistemi di supporto alle decisioni". *Etica e tecnologia dell'informazione*, 12(1):29–42.
- ⁸¹ Obermeyer, Ziad, Powers, Brian, Vogeli, Christine e Mullainathan, Sendhil. 2019. "Dissezione dei pregiudizi razziali in un algoritmo utilizzato per gestire la salute delle popolazioni". *Scienza*, 366 (6464): 447–453.
- ⁸² Schauer, Federico. 2006. Profili, probabilità e stereotipi. Harvard University Press, Cambridge, MA.
- 83 Lippert-Rasmussen, Kasper. 2011. "“Siamo tutti diversi”: discriminazione statistica e diritto essere trattato come un individuo". *Il giornale di etica*, 15 (1-2): 47–59.
- 84 Mitchell, Tom M.. 1980. "La necessità di pregiudizi nell'apprendimento delle generalizzazioni". Rapporto tecnico, Dipartimento di Informatica, Laboratorio per la Ricerca in Informatica, Rutgers University, New Brunswick, NJ.
- 85 Oreskes, Naomi, Shrader-Frechette, Kristin e Belitz, Kenneth. 1994. "Verifica, validazione e conferma di modelli numerici nelle scienze della terra". *Scienza*, 263 (5147): 641–646.
- ⁸⁶ Malik, Momin M. 2020. "Una gerarchia di limitazioni nell'apprendimento automatico". arXiv prestampa *arXiv:2002.05193*.
- 87 Mayer-Schönberger, Viktor e Cukier, Kenneth. 2013. Big Data: una rivoluzione che trasformerà il modo in cui viviamo, lavoriamo e pensiamo. Harper Business, New York, New York.
- ⁸⁸ Pasquale, Frank. 2018. "Quando il machine learning è faccialemente invalido". *Comunicazioni dell'ACM*, 61(9):25–27.
- 89 Kim, Pauline T. e Hanson, Erika. 2016. "Analisi delle persone e regolamentazione delle informazioni ai sensi del Fair Credit Reporting Act". *Giornale di diritto della Saint Louis University*, 61 (1): 17–34.
- 90 Salganik, Matthew J, Lundberg, Ian, Kindel, Alexander T, Ahearn, Caitlin E, Al-Ghoneim, Khaled, Almaatouq, Abdullah, Altschul, Drew M, Brand, Jennie E, Carnegie, Nicole Bohme, Compton, Ryan James, et al.. 2020. "Misurare la prevedibilità dei risultati della vita con una collaborazione scientifica di massa". Atti dell'Accademia Nazionale delle Scienze, 117(15):8398–8403.
- 91 Chouldechova, Alexandra, Putnam-Hornstein, Emily, Benavides-Prado, Diana, Fialko, Oleksandr e Vaithianathan, Rhema. 2017. "Un caso di studio sul processo decisionale assistito da algoritmi nelle decisioni di screening della hotline per il maltrattamento sui minori". In Atti di ricerca sull'apprendimento automatico, volume 81, pagine 1–15.
- 92 Huq, Aziz Z.. 2020. "Un diritto a una decisione umana". *Virginia Law Review*, 106(3):611–688.
- 93 Ustun, Berk, Spangher, Alexander e Liu, Yang. 2019. "Ricorso impugnabile nella classificazione lineare". In Conferenza su equità, responsabilità e trasparenza, pagine 10–19.
- 94 Milli, Smitha, Miller, John, Dragan, Anca D. e Hardt, Moritz. 2019. "Il costo sociale della classificazione strategica". In Conferenza su equità, responsabilità e trasparenza, pagine 230–239.
- 95 Hu, Lily, Immorlica, Nicole e Vaughan, Jennifer Wortman. 2019. "Gli effetti disparati della manipolazione strategica". In Conferenza su equità, responsabilità e trasparenza, pagine 259–268.
- 96 Karimi, Amir-Hossein, Barthe, Gilles, Schölkopf, Bernhard, e Valera, Isabel. 2022. "Un'indagine sul ricorso algoritmico: spiegazioni contrastive e raccomandazioni consequenziali". Sondaggi informatici ACM .

- 97 Kiviat, Barbara. 2019. "I limiti morali delle pratiche predittive: il caso dell'assicurazione basata sul credito punteggi". *American Sociological Review*, 84(6):1134–1158.
- 98 Miller, John, Milli, Smitha e Hardt, Moritz. 2020. "La classificazione strategica è la modellazione causale travestimento". *Negli Atti della 37a Conferenza Internazionale sull'Apprendimento Automatico*, pagine 6917–6926.
- 99 O'Neil, Cathy. 2017. *Armi di distruzione della matematica*. Penguin Random House, New York, NY.
- ¹⁰⁰ Bambauer, Jane e Zarsky, Tal. 2018. "Il gioco degli algoritmi". *Notre Dame L. Rev.*, 94:1.
- ¹⁰¹ Chen, Irene Y., III, Hal Daumé e Barocas, Solon. 2021. "I tanti ruoli che il ragionamento causale gioca nel ragionamento sull'equità nell'apprendimento automatico". *Nel workshop NeurIPS sull'algoritmo Equità attraverso la lente della causalità e della robustezza*.
- ¹⁰² Desrosières, Alain. 1998. *La politica dei grandi numeri: una storia del ragionamento statistico*. Harvard Stampa universitaria.
- 103 Porter, Theodore M. 2020. *L'ascesa del pensiero statistico, 1820-1900*. Stampa dell'Università di Princeton.
- 104 Bouk, Dan. 2015. *Come i nostri giorni sono diventati contati: il rischio e l'ascesa dell'individuo statistico*. Stampa dell'Università di Chicago.
- 105 Crenshaw, Kimberlé W. 2017. *Sull'intersezionalità: scritti essenziali*. La nuova stampa.
- ¹⁰⁶ Pipeline, Ryan, Varadarajan, Avinash V, Blumer, Katy, Liu, Yun, McConnell, Michael V, Corrado, Greg S, Peng, Lily e Webster, Dale R. 2018. "Previsione dei fattori di rischio cardiovascolare da fotografie del fondo retinale tramite deep learning". *Natura Ingegneria Biomedica*, 2(3):158–164.
- 107 Feldman, Michael, Friedler, Sorelle A, Moeller, John, Scheidegger, Carlos e Venkatasubramanian, Suresh. 2015. "Certificare e rimuovere gli impatti disparati". *Nell'Proc. 21° SIGKDD*. ACM.
- ¹⁰⁸ Ryan, Michelle K e Haslam, S Alexander. 2005. "La scogliera di vetro: la prova che le donne lo sono sovrarappresentati in posizioni di leadership precarie". *British Journal of Management*, 16(2):81–90.
- 109 Cook, Alison e Glass, Christy. 2014. "Sopra il soffitto di vetro: quando sono le donne e la razza/etnia minoranze promosse ad amministratore delegato?" *Giornale di gestione strategica*, 35(7):1080–1089.
- ¹¹⁰ Platt, John et al.. 1999. "Output probabilistici per macchine a vettori di supporto e confronti con metodi di verosimiglianza regolarizzata". *Progressi nei classificatori a margine ampio*, 10(3):61–74.
- ¹¹¹ Ding, Frances, Hardt, Moritz, Miller, John e Schmidt, Ludwig. 2021. "Adulto in pensione: Novità set di dati per un apprendimento automatico equo". *Progressi nei sistemi di elaborazione delle informazioni neurali*, 34.
- ¹¹² Liu, Lydia T, Simchowitz, Max e Hardt, Moritz. 2019. "Il criterio di equità implicita del apprendimento non vincolato". *Nella conferenza internazionale sull'apprendimento automatico*, pagine 4051–4060. PMLR.
- 113 Copertina, Thomas M. 1999. *Elementi di teoria dell'informazione*. John Wiley & Figli.
- 114 Hardt, Moritz, Price, Eric e Srebro, Nati. 2016. "Pari opportunità nell'apprendimento supervisionato". In *Advances in Neural Information Processing Systems*, pagine 3315–3323.
- 115 Wassermann, Larry. 2010. *Tutte le statistiche: un corso conciso sull'inferenza statistica*. Springer.
- 116 Il Consiglio della Federal Reserve. 2007. "Relazione al congresso sul credit scoring e i suoi effetti sulla disponibilità e accessibilità del credito". <https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/>. Accesso: 29-05-2018.
- 117 Hacking, Ian. 1990. *L'addomesticamento del caso*. Stampa dell'Università di Cambridge.
- ¹¹⁸ —. 2006. *L'emergere della probabilità: uno studio filosofico delle prime idee sulla probabilità, induzione e inferenza statistica*. Stampa dell'Università di Cambridge.
- 119 Hardt, Moritz e Recht, Benjamin. 2022. *Modelli, previsioni e azioni: Fondamenti della macchina apprendimento*. Stampa dell'Università di Princeton.
- ¹²⁰ Hutchinson, Ben e Mitchell, Margaret. 2019. "50 anni di (non) equità dei test: lezioni per le macchine apprendimento". In *Conferenza su equità, responsabilità e trasparenza*, pagine 49–58.
- ¹²¹ Chiara, T. Anne. 1966. "Bias del test: validità del test attitudinale scolastico per negri e bianchi studenti degli istituti integrati". *Serie di bollettini di ricerca ETS*, 1966(2):i–23.
- ¹²² —. 1968. "Distorsione del test: previsione dei voti degli studenti neri e bianchi nelle università integrate". *Giornale di misurazione educativa*, 5(2):115–124.
- 123 Darlington, Richard B. 1971. "Un altro sguardo all'equità culturale". *Journal of Educational Measurement*, 8(2):71–82.
- 124 Einhorn, Hillel J e Bass, Alan R. 1971. "Considerazioni metodologiche rilevanti per la discriminazione nei test di occupazione." *Bollettino psicologico*, 75(4):261.
- 125 Thorndike, Robert L. 1971. "Concetti di equità culturale". *Giornale di misurazione educativa*, 8(2):63–70.

- ¹²⁶ Lewis, Mary A. 1978. "Un confronto tra tre modelli per determinare l'equità del test". Rapporto tecnico , Federal Aviation Administration Washington DC Office of Aviation Medicine.
- 127 Calders, Toon, Kamiran, Faisal e Pechenizkiy, Mykola. 2009. "Costruire classificatori con vincoli di indipendenza". Nel Proc. IEEE ICDMW, pagine 13–18.
- ¹²⁸ Kamiran, Faisal e Calders, Toon. 2009. "Classificare senza discriminare". Nell'Proc. 2a Conferenza Internazionale su Computer, Controllo e Comunicazione.
- 129 Zemel, Richard S., Wu, Yu, Swersky, Kevin, Pitassi, Toniann e Dwork, Cynthia. 2013. "Apprendimento rappresentazioni giuste". Nella conferenza internazionale sull'apprendimento automatico.
- 130 Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer e Zemel, Richard. 2012. "Equità attraverso la consapevolezza". Nell'Proc. 3° ITCS, pagine 214–226.
- 131 Zafar, Muhammad Bilal, Valera, Isabel, Gómez Rodriguez, Manuel e Gummadi, Krishna P.. 2017. "Equità oltre trattamenti disparati e impatti disparati: classificazione dell'apprendimento senza maltrattamenti disparati". Nell'Proc. 26esimo WWW.
- 132 Woodworth, Blake E., Gunasekar, Suriya, Ohannessian, Mesrob I., e Srebro, Nathan. 2017. "Apprendimento di predittori non discriminatori". Nell'Proc. 30° COLT, pagine 1920–1953.
- 133 Angwin, Julia, Larson, Jeff, Mattu, Surya e Kirchner, Lauren. 2016. "Pregiudizio della macchina". ProPublica.
- 134 Dieterich, William, Mendoza, Christina e Brennan, Tim. 2016. "Scale di rischio Compas: dimostrare accuratezza, equità e parità predittiva".
- 135 Berk, Richard, Heidari, Hoda, Jabbari, Shahin, Kearns, Michael e Roth, Aaron. 2017. "L'equità nella valutazione dei rischi della giustizia penale: lo stato dell'arte". Stampe elettroniche ArXiv, 1703.09207.
- 136 Chouldechova, Alexandra. 2016. "Previsione corretta con impatti disparati: uno studio sui bias negli strumenti di previsione della recidività". Nel workshop su equità, responsabilità e trasparenza nell'apprendimento automatico.
- 137 Kleinberg, Jon M., Mullainathan, Sendhil, e Raghavan, Manish. 2017. "Compromessi intrinseci nel l'equa determinazione dei punteggi di rischio". Proc. 8° ITC.
- 138 Pleiss, Geoff, Raghavan, Manish, Wu, Felix, Kleinberg, Jon e Weinberger, Kilian Q. 2017. "On equità e calibrazione". Nei progressi nei sistemi di elaborazione delle informazioni neurali.
- 139 Hellman, Debora. 2007. Quando la discriminazione è sbagliata? Harvard University Press, Cambridge, MA.
- 140 Lippert-Rasmussen, Kasper. 2013. Nati liberi e uguali? Un'indagine filosofica sulla natura di Discriminazione. Oxford University Press, Oxford, Regno Unito.
- 141 Sunstein, Cass R. 1994. "Il principio dell'anticasta". Michigan Law Review, 92(8):2410.
- 142 Cantante, Pietro. 1978. "La discriminazione razziale è arbitraria?" Filosofia, 8 (2-3): 185–203.
- 143 Alessandro, Larry. 1992. "Cosa rende sbagliata la discriminazione ingiusta? Bias, preferenze, Stereotipi e proxy". Revisione giuridica dell'Università della Pennsylvania, 141(1):149.
- 144 Arneson, Richard J.. 2006. "Che cos'è la discriminazione ingiusta". San Diego Law Review, 43(4):775–808.
- 145 Eidelson, Benjamin. 2015. Discriminazione e mancanza di rispetto. Oxford University Press, New York, New York.
- 146 Balkin, J M. 1997. "La Costituzione dello status". The Yale Law Journal, 106(8):2313–2374.
- 147 Hoffman, Sharona. 2011. "L'importanza dell'immutabilità nella legislazione sulla discriminazione sul lavoro". William & Mary Law Review, 52(5):1483–1546.
- 148 Clarke, Jessica A.. 2015. "Contro l'immutabilità". Yale Law Journal, 125(1):1–102.
- 149 Hellman, Debora. 2020. "La discriminazione indiretta e il dovere di evitare di aggravare l'ingiustizia". In Hugh Collins e Tarunabh Khaitan (a cura di), Fondamenti della legge sulla discriminazione indiretta. Bloomsbury Publishing, Londra.
- 150 Schauer, Federico. 2017. "Discriminazione statistica (e non statistica)". In Kasper Lippert- Rasmussen (a cura di), The Routledge Handbook of the Ethics of Discrimination. Routledge, New York, NY.
- 151 Hellman, Debora. 2017. "Discriminazione e significato sociale". In Kasper Lippert-Rasmussen (a cura di), Il manuale Routledge dell'etica della discriminazione. Routledge, New York, NY.
- 152 Prince, Anya ER e Schwarcz, Daniel. 2020. "La discriminazione per procura nell'era dell'intelligenza artificiale e dei Big Data". Iowa Law Review, 105(3):1257–1318.
- 153 Anderson, Elizabeth S. 1999. "Qual è il punto dell'uguaglianza?" Etica, 109 (2): 287–337.
- 154 Rawls, John. 1998. Una teoria della giustizia. Harvard University Press, Cambridge, MA.

- 155 Arneson, Richard. 2018. "Quattro concezioni di pari opportunità". *Il Giornale Economico*, 128(612):F152–F173.
- 156 Fishkin, Giuseppe. 2013. *Colli di bottiglia*. Oxford University Press, New York, New York.
- 157 Roemer, John. 2000. *Pari opportunità*. Harvard University Press, Cambridge, MA.
- 158 Selbst, Andrew D., Boyd, Danah, Friedler, Sorelle A., Venkatasubramanian, Suresh e Vertesi, Janet. 2019. "Equità e astrazione nei sistemi sociotecnici". In *Conferenza su equità, responsabilità e trasparenza*, pagine 59–68.
- 159 Liu, Lydia T., Dean, Sarah, Rolf, Esther, Simchowitz, Max e Hardt, Moritz. 2017. "Impatto ritardato del Fair Machine Learning". Negli Atti della 35a Conferenza Internazionale sull'Apprendimento Automatico , volume 80, pagine 3150–3158.
- 160 Phillips, Anna. 2004. "Difendere l'uguaglianza dei risultati". *Giornale di filosofia politica*, 12 (1): 1–19.
- 161 Rieke, Aaron e Koepke, Logan. 2015. "Portati fuori strada: lead generation online e prestiti con anticipo sullo stipendio". Rapporto tecnico, Uptown, Washington, DC.
- 162 Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie H e Roth, Aaron. 2016. "Equità nell'apprendimento : banditi classici e contestuali". Nei progressi nei sistemi di elaborazione delle informazioni neurali, pagine 325–333.
- 163 Perry, Ronen e Zarsky, Tal. 2015. "'Che le probabilità siano sempre a tuo favore": lotterie in diritto". *Alabama Law Review*, 66(5):1035–1098.
- 164 Creel, Kathleen e Hellman, Deborah. 2022. "Il Leviatano algoritmico: arbitrarietà, equità e opportunità nei sistemi decisionali algoritmici". *Giornale canadese di filosofia*, pagine 1–18.
- 165 Matwin, Stan, Yu, Shipeng, Farooq, Faisal, Corbett-Davies, Sam, Pierson, Emma, Feller, Avi, Goel, Sharad e Huq, Aziz. 2017. "Il processo decisionale algoritmico e il costo dell'equità". Conferenza internazionale sulla scoperta della conoscenza e sul data mining, pagine 797–806.
- 166 Hellmann, Debora. 2022. "Misurazione dell'equità algoritmica". *Virginia Law Review*, 106(4):811–866.
- 167 Krishnapriya, KS, Vangara, Kushal, King, Michael C., Albiero, Vitor, e Bowyer, Kevin. 2019. "Caratterizzazione della variabilità nell'accuratezza del riconoscimento facciale rispetto alla razza". volume 00 della conferenza IEEE/CVF 2019 sui workshop sulla visione artificiale e il riconoscimento dei modelli (CVPRW), pagine 2278–2285.
- 168 Blodgett, Su Lin, Green, Lisa e O'Connor, Brendan. 2016. "Variazione dialettale demografica nei social media: un caso di studio sull'inglese afro-americano". In Conferenza sui metodi empirici nell'elaborazione del linguaggio naturale, pagine 1119–1130.
- 169 Huq, Aziz Z. 2018. "Equità razziale nella giustizia penale algoritmica". *Duca LJ*, 68:1043.
- 170 Passi, Samir e Barocas, Solone. 2019. "Formulazione dei problemi ed equità". Nella Conferenza del Equità, responsabilità e trasparenza, pagine 39–48.
- 171 Black, Emily, Raghavan, Manish e Barocas, Solon. 2022. "Molteplicità dei modelli: opportunità, preoccupazioni e soluzioni". In Conferenza su equità, responsabilità e trasparenza, pagine 850–863.
- 172 Rodolfa, Kit T., Lamba, Hemank, e Ghani, Rayid. 2021. "Osservazione empirica di compromessi trascurabili tra equità e accuratezza nell'apprendimento automatico per le politiche pubbliche". *Natura Macchina Intelligenza*, 3(10):896–904.
- 173 Friedler, Sorelle A., Scheidegger, Carlos e Venkatasubramanian, Suresh. 2021. "La (Im)possibilità dell'equità". *Comunicazioni dell'ACM*, 64(4):136–143.
- 174 Hannah-Jones, Nikole. 2020. "Cosa è dovuto". *Il New York Times*.
- 175 Bickel, Peter J, Hammel, Eugene A, O'Connell, J William, et al.. 1975. "Sex bias in graduate admissions: Data from Berkeley". *Scienza*, 187(4175):398–404.
- 176 Humphrey, Linda L., Chan, Benjamin KS e Sox, Harold C.. 2002. "Terapia ormonale sostitutiva postmenopausale e prevenzione primaria delle malattie cardiovascolari". *Annali di medicina interna*, 137(4):273–284.
- 177 Berkson, Joseph. 2014. "Limiti dell'applicazione dell'analisi quadrupla delle tabelle ai dati ospedalieri". *Giornale internazionale di epidemiologia*, 43(2):511–515. Ristampare.
- 178 Moneta-Koehler, Liane, Brown, Abigail M., Petrie, Kimberly A., Evans, Brent J. e Chalkley, Roger. 2017. "I limiti del gre nel predire il successo nella scuola di specializzazione in biomedicina". *PLOS UNO*, 12(1):1–17.
- 179 Hall, Joshua D., O'Connell, Anna B. e Cook, Jeanette G.. 2017. "Predictors of student produt-attività nelle applicazioni delle scuole di specializzazione in biomedicina". *PLOS UNO*, 12(1):1–14.

- ¹⁸⁰ Perla, Giudea. 2009. Causalità. Stampa dell'Università di Cambridge.
- ¹⁸¹ Deaton, Angus e Cartwright, Nancy. 2018. "Comprensione e incomprensione degli studi randomizzati e controllati". *Scienze sociali e medicina*, 210: 2–21.
- ¹⁸² Pearl, Giudea e Mackenzie, Dana. 2018. *Il libro dei perché: la nuova scienza di causa ed effetto. Libri di base.*
- 183 Glymour, M Maria. 2006. "Utilizzo di diagrammi causali per comprendere i problemi comuni nell'epidemiologia sociale ". *Metodi di epidemiologia sociale*, pagine 393–428.
- 184 Krieger, Nancy. 2011. "Epidemiologia e salute delle persone: teoria e contesto".
- 185 Peters, Jonas, Janzing, Dominik e Schölkopf, Bernhard. 2017. *Elementi di inferenza causale*. MIT Premere.
- ¹⁸⁶ Baron, Reuben M e Kenny, David A. 1986. "La distinzione della variabile moderatore-mediatore nella ricerca psicologica sociale: considerazioni concettuali, strategiche e statistiche". *Giornale di personalità e psicologia sociale*, 51(6):1173.
- 187 Kusner, Matt J., Loftus, Joshua R., Russell, Chris e Silva, Ricardo. 2017. "Controfattuale equità". In *Advances in Neural Information Processing Systems*, pagine 4069–4079.
- ¹⁸⁸ Pearl, Judea, Glymour, Madelyn e Jewell, Nicholas P.. 2016. *Inferenza causale in statistica: un primer*. Wiley.
- 189 Egan, Patrick J. 2020. "Identità come variabile dipendente: come gli americani spostano le loro identità verso alinearsi con la loro politica". *Giornale americano di scienze politiche*, 64(3):699–716.
- 190 Glasgow, Joshua, Haslanger, Sally, Jeffers, Chike e Spencer, Quayshawn. 2019. "Cos'è la razza?: Quattro visioni filosofiche".
- 191 Bowker, Geoffrey C e Star, Susan Leigh. 2000. *Sistemare le cose: classificazione e sue conseguenze*. Stampa del MIT.
- 192 Fields, Karen E. e Fields, Barbara J.. 2014. *Racecraft: L'anima della disegualanza nella vita americana*. Verso.
- 193 Beniamino, Ruha. 2019. *Corsa dopo la tecnologia*. Politica.
- 194 Hacking, Ian. 2000. La costruzione sociale di cosa? Stampa dell'Università di Harvard.
- 195 Haslanger, Sally. 2012. *Resistere alla realtà: costruzione sociale e critica sociale*. Università di Oxford Premere.
- 196 Mallón, Ron. 2018. *La costruzione dei generi umani*. Stampa dell'Università di Oxford.
- 197 Cartwright, Nancy. 2006. *Cacciare le cause e anche usarle*. Stampa dell'Università di Cambridge.
- 198 Hacking, Ian. 2006. "Inventare le persone". *London Review of Books*, 28(16).
- 199 Holland, Paul W.. 1986. "Statistica e inferenza causale". *Giornale dell'American Statistical Association (JASA)*, 81: 945–970.
- ²⁰⁰ VanderWeele, Tyler J. e Robinson, Whitney R.. 2014. "Sull'interpretazione causale della razza nelle regressioni che si adattano alle variabili confondenti e mediatici". *Epidemiologia*.
- ²⁰¹ Greiner, D. James e Rubin, Donald B.. 2011. "Effetti causali di caratteristiche immutabili percepite ". La revisione di economia e statistica, 93 (3): 775–785.
- ²⁰² Kohler-Hausmann, Issa. 2019. "Eddie Murphy e i pericoli del pensiero causale controfattuale sull'individuazione della discriminazione razziale". SSRN.
- 203 Spirtes, Peter, Glymour, Clark N, Scheines, Richard, Heckerman, David, Meek, Christopher, Cooper, Gregory e Richardson, Thomas. 2000. *Causa, previsione e ricerca*. Stampa del MIT.
- 204 Morgan, Stephen L. e Winship, Christopher. 2014. *Controfattuali e inferenza causale*. Campania University Press.
- 205 Imbens, Guido W. e Rubin, Donald B.. 2015. *Inferenza causale per la statistica, la società e la biomedicina Scienze*. Stampa dell'Università di Cambridge.
- ²⁰⁶ Angrist, Joshua D. e Jörn-Steffen, Pischke. 2009. *Econometria per lo più innocua: il compagno di un empirista*. Stampa dell'Università di Princeton.
- 207 Hernán, Miguel e Robins, James. 2019. *Inferenza causale*. Boca Raton: Chapman & Hall/CRC, imminente.
- ²⁰⁸ Simpson, Edward H. 1951. "L'interpretazione dell'interazione nelle tabelle di contingenza". *Giornale della Royal Statistical Society: Serie B (metodologica)*, 13(2):238–241.
- 209 Hernán, Miguel A, Clayton, David, e Keiding, Niels. 2011. "Il paradosso dei Simpson svelato". *Giornale internazionale di epidemiologia*, 40(3):780–785.
- ²¹⁰ Zhang, Lu, Wu, Yongkai e Wu, Xintao. 2017. "Un quadro causale per scoprire e rimuovere la discriminazione diretta e indiretta". Nell'Proc. 26° IJCAI, pagine 3929–3935.

- ²¹¹ Russell, Chris, Kusner, Matt J., Loftus, Joshua R. e Silva, Ricardo. 2017. "Quando i mondi si scontrano: integrare diverse ipotesi controfattuali nell'equità". In Advances in Neural Information Processing Systems, pagine 6417–6426.
- ²¹² Chiappa, Silvia. 2019. "Equità controfattuale specifica del percorso". Nell'Proc. 33esimo AAAI, volume 33, pagine 7801–7808.
- 213 Kilbertus, Niki, Rojas-Carulla, Mateo, Parascandolo, Giambattista, Hardt, Moritz, Janzing, Dominik , e Schölkopf, Bernhard. 2017. "Evitare la discriminazione attraverso il ragionamento causale". In Advances in Neural Information Processing Systems, pagine 656–666.
- 214 Nabi, Razieh e Shpitser, Ilya. 2018. "Equa inferenza sui risultati". Nell'Proc. 32a AAAI, pagine 1931-1940.
- 215 Chiappa, Silvia e Isaac, William S.. 2019. "A causal bayesian network viewpoint on fairness". arxiv.org, arXiv:1907.06430.
- ²¹⁶ Kasirzadeh, Atoosa e Smart, Andrew. 2021. "L'uso e l'abuso dei controfattuali nell'apprendimento automatico etico". In Conferenza su equità, responsabilità e trasparenza, pagine 228–236.
- 217 Krieger, Nancy. 2014. "Sull'interpretazione causale della razza". Epidemiologia, 25(6):937.
- ²¹⁸ —. 2014. "Discriminazioni e disuguaglianze sanitarie". Giornale internazionale dei servizi sanitari, 44(4):643–710 . 219 . 1896. "Plessy contro Ferguson".
- ²²⁰ Wilkerson, Isabel. 2011. Il calore di altri soli: l'epica storia della grande migrazione americana. Penguin Random House, New York, NY.
- ²²¹ Keele, Luke, Cubbison, William e White, Ismail. 2021. "Soppressione dei voti neri: un caso di studio storico sulle restrizioni al voto in Louisiana". Revisione americana di scienze politiche, 115(2):694–700.
- ²²² Rothstein, Richard. 2018. Il colore della legge: una storia dimenticata di come il nostro governo ha segregato l'America. WW Norton & Company, New York, NY.
- 223 Walker, Giulietta EK. 1998. Storia del business nero in America: capitalismo, razza, imprenditorialità. Evoluzione del business moderno. Twayne Editori, New York, NY.
- 224 Vaas, Francis J. 1965. "Titolo vii: Storia legislativa". aC Indo. & Com. L. Rev., 7:431.
- 225 Friedan, Betty. 2001. La mistica femminile. WW Norton & Company, New York, NY.
- ²²⁶ . 1973. "Roe contro Wade".
- ²²⁷ . 1976. "Rapporto Senato 589 del 94° congresso".
- ²²⁸ Eskridge, William N.. 2008. Passioni disonorevoli: leggi sulla sodomia in America, 1861-2003. Vichingo, New York, New York.
- ²²⁹ . 2003. "Lawrence contro Texas".
- ²³⁰ . 2015. "Obergefell contro Hodges".
- ²³¹ . 2020. "Bostock contro contea di Clayton, Georgia".
- 232 Okoro, Catherine A., Hollis, NaTasha D., Cyrus, Alissa C. e Griffin-Blake, Shannon. 2018. "Prevalenza delle disabilità e accesso all'assistenza sanitaria in base allo stato e al tipo di disabilità tra gli adulti – Stati Uniti, 2016". Rapporto settimanale sulla morbilità e mortalità, 67(32):882–887.
- 233 Imbottito, Carol e Humphries, Tom. 2006. All'interno della cultura dei sordi. Stampa dell'Università di Harvard, Cambridge, MA.
- 234 Fleischer, Doris Zames e Zames, Frieda. 2013. Il movimento per i diritti dei disabili: dalla carità al confronto. JSTOR.
- ²³⁵ . 2019. "Facebook, inc."
- 236 Rosenblat, Alex, Levy, Karen EC, Barocas, Solon e Hwang, Tim. 2016. "Gusti discriminanti: Le valutazioni dei clienti come veicoli di pregiudizi". Dati e società, pagine 1–21.
- 237 Lucas, Lauren Sudeall. 2015. "Identità come procura". Columbia Law Review, 115(6):1605–1674. 238 . 1964. "Cuore di Atlanta Motel, Inc. contro Stati Uniti".
- 239 Dau-Schmidt, Kenneth Glenn e Sherman, Ryland. 2013. "L'occupazione e il progresso economico degli afroamericani nel ventesimo secolo". Jindal Journal of Public Policy, 1(2):95–116.
- 240 Walker, Christopher J.. 2018. "Attaccare Auer e Chevron Deference: una revisione della letteratura". IL Georgetown Journal of Law & Public Policy, 16(1):103–122.
- ²⁴¹ . 1941. "Ordinanza dirigenziale n. 8802".
- ²⁴² . 2015. "Domande e risposte sulla guida applicativa dell'EEOC sulla discriminazione in gravidanza e questioni correlate".

- 243 Valentin, Iram. 1997. "Titolo IX: Una breve storia". Relazione tecnica, Equity Resource Center, Newton, MA.
- 244 Graham , Hugh Davis. 1998. "La tempesta sul Grove City College: regolamentazione dei diritti civili, istruzione superiore e amministrazione Reagan". Storia dell'educazione trimestrale, 38(4):407–429.
- 245 Melnick, R. Shep. 2020. "Analisi delle norme finali del Titolo IX del Dipartimento dell'Istruzione in materia sessuale cattiva condotta". Rapporto tecnico, The Brookings Institution, Washington, DC.
- 246 della Pubblica Istruzione, Dipartimento. 2021. "Applicazione del titolo ix degli emendamenti sull'istruzione del 1972 rispetto alla discriminazione basata sull'orientamento sessuale e sull'identità di genere alla luce del caso Bostock v. Clayton County".
- 247 Katz, Martin J. 2008. "Unificare trattamenti disparati (davvero)". Giornale legale di Hastings, 59(3):643.
- 248 Price, Robert N.. 2016. "Griggs v. Duke Power Co.: Il primo punto di riferimento ai sensi del Titolo VII del Civil Rights Act del 1964". Giornale giuridico del sud-ovest, 25(3):484–493.
- 249 Occhio, Katie. 2021. "La teoria dei ma-per del diritto antidiscriminatorio". Va. L. Rev., 107:1621.
- 250 . 2021. "Manuale giuridico del Titolo VI". Rapporto tecnico, Divisione per i diritti civili Dipartimento di Giustizia degli Stati Uniti, Washington, DC.
- 251 . 1963. "Rapporto della Camera dei Rappresentanti n. 914 Pt 2, 88° Congresso, 1a Sessione".
- 252 . 2015. "Dipartimento del Texas per l'edilizia abitativa e gli affari comunitari contro progetto di comunità inclusive, inc."
- 253 Porter, Nicole Buonocore. 2019. "Un nuovo sguardo alla difesa del disagio indebito dell'ada". Mo. L. Rev., 84:121.
- 254 Menand, Louis. 2020. "Il significato mitevole dell'azione affermativa". Il New Yorker.
- 255 Bagenstos, Samuel R.. 2014. "Formalismo e responsabilità del datore di lavoro ai sensi del Titolo VII". Università di Forum legale di Chicago, 2014(1):145–176.
- 256 Cairns, John W.. 1976. "La parità di credito arriva alle donne: un'analisi dell'Equal Credit Legge sulle opportunità". Rassegna legale di San Diego, 13(4):960–977.
- 257 Chay, Kenneth Y.. 1998. "L'impatto della politica federale sui diritti civili sul progresso economico dei neri: prove dalla legge sulle pari opportunità di lavoro del 1972". Revisione delle relazioni industriali e lavorative , 51(4):608–632. 258 . 2020. "Principi dei diritti civili per l'assunzione di tecnologie di valutazione".
- 259 Moss, Scott A.. 2007. "Lotta alla discriminazione mentre si combatte il contenzioso: una storia di due corti supreme". Fordham Law Review, 76(2):981–1013.
- 260 Sperino, Sandra F. e Thomas, Suja A.. 2017. *Unequal: How America's Courts Undermine Discrimination Law*. Oxford University Press;, New York, NY.
- 261 Ajunwa, Ifeoma. 2020. "Il paradosso dell'automazione come intervento anti-bias". Cardozo Law Review, 41(5):1671–1742.
- 262 Selmi, Michael. 2001. "Perché i casi di discriminazione sul lavoro sono così difficili da vincere?" Louisiana Law Review, 61(3):555–575.
- 263 Eyer, Katie R.. 2012. "Questa non è discriminazione: le credenze americane e i limiti dell'anti-legge sulla discriminazione". Minnesota Law Review, 96: 1275–1362.
- 264 Halpern, Stephen C.. 1995. *Sui limiti della legge: l'eredità ironica del titolo VI della legge civile del 1964 Legge sui diritti*. Johns Hopkins University Press, Baltimora, MD.
- 265 Edelman, Lauren B, Krieger, Linda H, Eliason, Scott R, Albiston, Catherine R e Mellema, Virginia. 2011. "Quando le organizzazioni governano: deferenza giudiziaria alle strutture occupazionali istituzionalizzate". Giornale americano di sociologia, 117(3):888–954.
- 266 . 1979. "Acciaierie contro Weber".
- 267 Kim, Paolina. 2022. "Algoritmi consapevoli della razza: equità, non discriminazione e azione affermativa". Revisione della legge della California.
- 268 Bent, Jason R.. 2020. "L'azione affermativa algoritmica è legale?" Il Georgetown Law Journal, 108(4):803–853. 269 . 1971. "Griggs contro Duke Power Co."
- 270 Kim, Pauline T.. 2017. "Discriminazione sul lavoro basata sui dati". Recensione di William & Mary Law, 58(3):857–936.
- 271 Raghavan, Manish, Barcas, Solon, Kleinberg, Jon, e Levy, Karen. 2020. "Mitigazione dei pregiudizi nelle assunzioni algoritmiche: valutazione di richieste e pratiche". In Conferenza su equità, responsabilità e trasparenza, pagine 469–481.

- 272 Duhigg, Charles. 2014. *Il potere dell'abitudine: perché facciamo quello che facciamo nella vita e negli affari*. Libro in brossura. Random House, New York, New York.
- 273 Landesberg, Martha K., Levin, Toby Milgrom, Curtin, Caroline G. e Lev, Ori. 1998. "Privacy online: un rapporto al Congresso". Rapporto tecnico, Federal Trade Commission, Washington, DC.
- 274 Ali, Muhammad, Sapiezynski, Piotr, Bogen, Miranda, Korolova, Aleksandra, Mislove, Alan, e Rieke, Aaron. 2019. "Discriminazione attraverso l'ottimizzazione: come la pubblicazione degli annunci di Facebook può portare a risultati distorti". Atti dell'ACM sull'interazione uomo-computer, 3 (CSCW):199.
- 275 Agan, Amanda e Starr, Sonja. 2017. "Ban the Box, precedenti penali e discriminazione razziale: un esperimento sul campo". Il giornale trimestrale di economia, 133(1):191–235.
- 276 Baroca, Solone e Nissenbaum, Elena. 2014. "La fine dei big data ruota attorno all'anomimato e al consenso". In Julia Lane, Victoria Stodden, Stefan Bender e Helen Nissenbaum (a cura di), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, pagine 44–75. Cambridge University Press, New York, New York.
- 277 Areheart, Bradley A. e Roberts, Jessica L.. 2019. "GINA, big data e il futuro della privacy dei dipendenti". Yale Law Journal, 128(3):710–790.
- 278 Strandburg, Katherine J.. 1999. "Regolamentazione e strumenti decisionali automatizzati imperscrutabili". Columbia Law Review, 119(7):1851–1886.
- 279 ___. 2021. "Giudicare con strumenti decisionali imperscrutabili". In Marcello Pelillo e Teresa Scantamburlo (a cura di), *Machines We Trust: Perspectives on Dependable AI*. MIT Press, Cambridge, MA.
- 280 Cedro, Danielle Keats. 2007. "Il giusto processo tecnologico". Washington UL Rev., 85:1249.
- 281 Selbst, Andrew D e Barocas, Solon. 2018. "Il fascino intuitivo delle macchine spiegabili". Fordham Law Review, 87(3):1085.
- 282 Andreou, Athanasios, Venkatadri, Giridhari, Goga, Oana, Gummadi, Krishna, Loiseau, Patrick e Mislove, Alan. 2018. "Indagare sui meccanismi di trasparenza degli annunci nei social media: un caso di studio sulle spiegazioni di Facebook". In NDSS 2018 - Simposio sulla sicurezza delle reti e dei sistemi distribuiti.
- 283 Kaminski, Margot E e Malgieri, Gianclaudio. 2020. "Valutazioni di impatto algoritmico ai sensi del GDPR: produrre spiegazioni a più livelli". Legge internazionale sulla privacy dei dati, 11(2):125–144.
- 284 Selbst, Andrew D.. 2021. "Una visione istituzionale delle valutazioni di impatto algoritmico". Harvard Journal of Law & Technology, 35(1):117–191.
- 285 Lovelace, Ada e DataKind, Regno Unito. 2020. "Esaminare la scatola nera: strumenti per la valutazione algoritmica sistemi". Relazione tecnica, Relazione tecnica, Ada Lovelace Institute.
- 286 Costanza-Chock, Sasha, Raji, Inioluwa Deborah e Buolamwini, Joy. 2022. "Chi controlla i revisori? raccomandazioni da una scansione sul campo dell'ecosistema di audit algoritmico". In Conferenza su equità, responsabilità e trasparenza, pagine 1571–1583.
- 287 Ajunwa, Ifeoma. 2021. "Un imperativo di audit per le assunzioni automatizzate". Harvard Journal of Law & Tecnologia, 34(2):621–299.
- 288 Sinclair, Upton. 2006. La giungla. Pinguino, New York, New York.
- 289 Hoofnagle, Chris Jay. 2016. *Legge e politica sulla privacy della Federal Trade Commission*. Cambridge University Press, New York, New York.
- 290 ___. 2020. "Dichiarazione del Commissario Rebecca Kelly Slaughter in materia di libertà Chevrolet, Inc. d/b/a Bronx Honda". Ufficio del commissario Rebecca Kelly Slaughter, Federal Trade Commission.
- 291 Selbst, Andrew D e Barocas, Solon. 2023. "Intelligenza artificiale ingiusta: come l'intervento della FTC può superare i limiti della legge sulla discriminazione". Revisione giuridica dell'Università della Pennsylvania, 171.
- 292 ___. 2021. "Decisione e ordine, In re Everalbum, Inc., File n. 192-3172". Commissione federale per il commercio.
- 293 Pitofsky, Robert. 2005. "Passato, presente e futuro dell'applicazione delle norme antitrust presso la Federal Trade Commission". La rivista giuridica dell'Università di Chicago, 72(1):209–227.
- 294 Jillson, Elisa. 2021. "Puntare alla verità, all'equità e all'equità nell'uso dell'intelligenza artificiale da parte della vostra azienda".
- 295 Slaughter, Rebecca Kelly, Kopec, Janice e Batal, Mohamad. 2020. "Algoritmi e giustizia economica: una tassonomia dei dati e un percorso da seguire per la commissione federale per il commercio". Yale JL & Tech., 23:1.
- 296 Richards, Neil e Hartzog, Woodrow. 2021. "Un dovere di fedeltà per la normativa sulla privacy". Washington Revisione giuridica dell'Università, 99(3):961–1021.
- 297 Bogen, Miranda e Rieke, Aaron. 2018. "Cercasi aiuto: un esame degli algoritmi di assunzione, dell'equità e dei pregiudizi". Relazione tecnica, Relazione tecnica, Ripresa.

- 298 Wienk, Ronald E, Reid, Clifford E., Simonson, John C. e Eggers, Frederick J.. 1979. "Misurare la discriminazione razziale nei mercati immobiliari americani: l'indagine sulle pratiche del mercato immobiliare".
- 299 Ayres, Ian e Siegelman, Peter. 1995. "Razza e discriminazione di genere nella contrattazione per una nuova auto". *L'American Economic Review*, pagine 304–321.
- 300 Freeman, Jonathan B, Penner, Andrew M, Saperstein, Aliya, Scheutz, Matthias e Ambady, Nalini. 2011. "Guardare la parte: i segnali di status sociale modellano la percezione della razza". *PloS uno*, 6(9):e25107.
- 301 Bertrand, Marianne e Mullainathan, Sendhil. 2004. "Emily e Greg sono più occupabili di Lakisha e Jamal? un esperimento sul campo sulla discriminazione nel mercato del lavoro". *Revisione economica americana* , 94(4):991-1013.
- 302 Cercapersone, Devah. 2007. "L'uso di esperimenti sul campo per studi sulla discriminazione lavorativa: contributi, critiche e indicazioni per il futuro". *Gli Annali dell'Accademia americana di scienze politiche e sociali*, 609 (1): 104–133.
- 303 Kohler-Hausmann, Issa. 2018. "Eddie Murphy e i pericoli del pensiero causale controfattuale sull'individuazione della discriminazione razziale". *Ora. Rev. UL*, 113:1163.
- 304 Bertrand, Marianne e Duflo, Esther. 2017. "Esperimenti sul campo sulla discriminazione". Nel *Manuale degli esperimenti sul campo economico*, volume 1, pagine 309–393. Elsevier.
- 305 Quillian, Lincoln, Pager, Devah, Hexel, Ole e Midtbøen, Arnfinn H. 2017. "La meta-analisi degli esperimenti sul campo non mostra alcun cambiamento nella discriminazione razziale nelle assunzioni nel tempo". *Atti dell'Accademia Nazionale delle Scienze*, 114(41):10870–10875.
- 306 Blank, Rebecca M. 1991. "Gli effetti della revisione in doppio cieco rispetto a quella in singolo cieco: prove sperimentali dalla revisione economica americana". *L'American Economic Review*, pagine 1041–1067.
- 307 Pischke, Jörn-Steffen. 2005. "Metodi empirici in economia applicata: appunti delle lezioni".
- 308 Bertrand, Marianne, Duflo, Esther e Mullainathan, Sendhil. 2004. "Quanto dovremmo fidarci delle stime delle differenze nelle differenze?" *La rivista trimestrale di economia*, 119(1):249–275.
- 309 Kang, Sonia K, DeCelles, Katherine A, Tilcsik, András e Jun, Sora. 2016. "Curriculum imbiancati: razza e auto-presentazione nel mercato del lavoro". *Scienza amministrativa trimestrale*, 61(3):469–502.
- 310 Eren, Ozkan e Mocan, Naci. 2018. "Giudici emotivi e minorenni sfortunati". *americano Rivista economica: Economia applicata*, 10(3):171–205.
- 311 Danziger, Shai, Levav, Jonathan, e Avnaim-Pesso, Liora. 2011. "Fattori estranei in sede giudiziaria decisioni". *Atti dell'Accademia Nazionale delle Scienze*, 108(17):6889–6892.
- 312 Lakens, Daniele. 2017. "Giudici incredibilmente affamati". <https://daniellakens.blogspot.com/2017/07/impossibilmente-affamati-giudici.html>.
- 313 Weinshall-Margel, Keren e Shapard, John. 2011. "Fattori trascurati nell'analisi della libertà condizionale decisioni". *Atti dell'Accademia Nazionale delle Scienze*, 108(42):E833–E833.
- 314 Norton, Elena. 2010. "La svolta post-razziale della Corte Suprema verso una comprensione a somma zero del uguaglianza". *Wm. & Mary L. Rev.*, 52:197.
- 315 Ayres, Ian. 2005. "Tre test per misurare impatti disparati ingiustificati nel trapianto di organi: il problema del bias della "variabile inclusa"". *Prospettive in biologia e medicina*, 48(1):68–S87.
- 316 Simoiu, Camelia, Corbett-Davies, Sam, e Goel, Sharad. 2017. "Il problema dell'inframarginalità nei test di esito per la discriminazione". *Gli Annali di statistica applicata*, 11(3):1193–1216.
- 317 Lakkaraju, Himabindu, Kleinberg, Jon, Leskovec, Jure, Ludwig, Jens e Mullainathan, Sendhil. 2017. "Il problema delle etichette selettive: valutazione delle previsioni algoritmiche in presenza di non osservabili". In *Conferenza internazionale sulla scoperta della conoscenza e sul data mining*, pagine 275–284. ACM.
- 318 Bird, Sarah, Barocas, Solon, Crawford, Kate, Diaz, Fernando e Wallach, Hanna. 2016. "Esplorare o sfruttare? implicazioni sociali ed etiche della sperimentazione autonoma nell'intelligenza artificiale". Nel *workshop su equità, responsabilità e trasparenza nell'apprendimento automatico*.
- 319 Becker, Gary S.. 1957. *L'economia della discriminazione*. Stampa dell'Università di Chicago.
- 320 Phelps, Edmund S.. 1972. "La teoria statistica del razzismo e del sessismo". *L'American Economic Review*, 62(4):659–661.
- 321 Arrow, Kenneth J.. 1973. "La teoria della discriminazione". In *Orley Ashenfelter e Albert Rees (a cura di), Discrimination in Labour Markets*, pagine 3–33. Stampa dell'Università di Princeton.
- 322 Agan, Amanda e Starr, Sonja. 2017. "Ban the box, precedenti penali e discriminazione razziale: un esperimento sul campo". *Il giornale trimestrale di economia*, 133(1):191–235.

- 323 Williams, Wendy M e Ceci, Stephen J. 2015. "Gli esperimenti di assunzione a livello nazionale rivelano una preferenza 2:1 tra i docenti per le donne con ruolo di ruolo". Atti dell'Accademia Nazionale delle Scienze, 112(17):5360–5365.
- 324 Neckerman, Kathryn M e Kirschenman, Joleen. 1991. "Strategie di assunzione, pregiudizi razziali e lavoratori dei centri urbani". Problemi sociali, 38 (4): 433–447.
- 325 Cercapersone, Devah e Pastore, Hana. 2008. "La sociologia della discriminazione: discriminazione razziale nei mercati del lavoro, dell'edilizia abitativa, del credito e dei consumi". Anna. Rev. Sociol, 34:181–209.
- 326 Rivera, Lauren A. 2016. Pedigree: come gli studenti d'élite ottengono lavori d'élite. Stampa dell'Università di Princeton.
- 327 Posselt, Julie R. 2016. Ammissioni all'interno dei laureati. Stampa dell'Università di Harvard.
- 328 Roth, Alvin E. 2003. "Le origini, la storia e il design della partita residente". Jama, 289(7):909–912.
- 329 Green, Lisa J. 2002. Inglese afroamericano: un'introduzione linguistica. Stampa dell'Università di Cambridge.
- 330 Dastin, Jeffrey. 2018. "Amazon elimina uno strumento segreto di reclutamento basato sull'intelligenza artificiale che mostrava pregiudizi contro le donne". Reuters.
- 331 Buranyi, Stefano. 2018. "Come convincere un robot a ottenere il lavoro".
- 332 De-Arteaga, Maria, Romanov, Alexey, Wallach, Hanna, Chayes, Jennifer, Borgs, Christian, Chouldechova , Alexandra, Geyik, Sahin, Kenthapadi, Krishnaram e Kalai, Adam Tauman. 2019. "Bias in bios: un caso di studio sul bias della rappresentazione semantica in un contesto ad alta posta in gioco". In Conferenza su equità, responsabilità e trasparenza, pagine 120–128. ACM.
- 333 Ramineni, Chaitanya e Williamson, David. 2018. "Comprensione delle differenze di punteggio medio tra il motore di punteggio automatizzato e-rater e gli esseri umani per gruppi basati su dati demografici nel test generale GRE". Serie di rapporti di ricerca ETS, 2018(1):1–31.
- 334 Amorim, Evelin, Cançado, Marcia, e Veloso, Adriano. 2018. "Punteggio automatizzato del saggio in presenza di valutazioni distorte". Nella conferenza del capitolo nordamericano dell'Associazione per la linguistica computazionale, pagine 229–237.
- 335 Sap, Maarten, Card, Dallas, Gabriel, Saadia, Choi, Yejin e Smith, Noah A. 2019. "Il rischio di pregiudizi razziali nel rilevamento del discorso d'odio". Nella riunione annuale dell'Associazione per la linguistica computazionale, pagine 1668–1678.
- 336 Kiritchenko, Svetlana e Mohammad, Saif. 2018. "Esaminare i pregiudizi di genere e di razza in duecento sistemi di analisi del sentimento". In Conferenza sulla semantica lessicale e computazionale, pagine 43–53. Associazione per la Linguistica Computazionale.
- 337 Tatman, Rachele. 2017. "Pregiudizi di genere e dialettali nei sottotitoli automatici di YouTube". Nel Workshop ACL sull'etica nell'elaborazione del linguaggio naturale, pagine 53–59. Associazione di Linguistica Computazionale , Valencia, Spagna.
- 338 Solaiman, Irene, Brundage, Miles, Clark, Jack, Askell, Amanda, Herbert-Voss, Ariel, Wu, Jeff, Radford, Alec e Wang, Jasmine. 2019. "Strategie di rilascio e impatti sociali dei modelli linguistici". arXiv prestampa [arXiv:1908.09203](https://arxiv.org/abs/1908.09203).
- 339 Buolamwini, Gioia e Gebru, Timnit. 2018. "Sfumature di genere: disparità di accuratezza intersezionale nella classificazione di genere commerciale". In Conferenza su equità, responsabilità e trasparenza, pagg 77–91.
- 340 de Vries, Terrance, Misra, Ishan, Wang, Changhan e van der Maaten, Laurens. 2019. " Il riconoscimento degli oggetti funziona per tutti?" Nella conferenza sui workshop sulla visione artificiale e sul riconoscimento dei modelli, pagine 52–59.
- 341 Shankar, Shreya, Halpern, Yoni, Breck, Eric, Atwood, James, Wilson, Jimbo e Sculley, D.. 2017. "Nessuna classificazione senza rappresentazione: valutazione delle questioni relative alla geodiversità in set di dati aperti per il mondo in via di sviluppo". Nel workshop NeurIPS 2017 : Machine Learning per il mondo in via di sviluppo.
- 342 Simonita, Tom. 2018. "Quando si parla di gorilla, google foto resta cieco". Wired, gennaio 13.
- 343 Hern, Alex. 2015. "Flickr deve affrontare lamentate per la codifica automatica 'offensiva' delle foto". Il Guardiano, 20.
- 344 Martineau, Parigi. 2019. del "Le città esaminano gli usi corretti e impropri <https://www.wired.com/story/facciale-corretto-improprio/>". www.wired.com/story/facciale-corretto-improprio/
- 345 O'Toole, Alice J, Deffenbacher, Kenneth, Abdi, Hervé e Bartlett, James C. 1991. "Simulare l'effetto dell'altra razza come un problema nell'apprendimento percettivo". Scienza della connessione, 3 (2): 163–178.
- 346 Frucci, Adamo. 2009. "Le webcam HP per il rilevamento del volto non riconoscono i neri". <https://gizmodo.com/hp-face-tracking-webcams-dont-recognize-black-people-5431190>.

- 347 McEntegart, Jane. 2010. "Kinect potrebbe avere problemi con gli utenti dalla pelle scura | la guida di Tom". <https://www.tomsguide.com/us/Microsoft-Kinect-Dark-Skin-Facial-Recognition,news-8638.html> .
- 348 Wilson, Benjamin, Hoffman, Judy e Morgenstern, Jamie. 2019. "Disuguaglianza predittiva nel rilevamento di oggetti". arXiv prestampa *arXiv:1902.11097*.
- 349 Turow, Joseph, King, Jennifer, Hoofnagle, Chris Jay, Bleakley, Amy e Hennessy, Michael. 2009. "Gli americani rifiutano la pubblicità personalizzata e le tre attività che la consentono". Disponibile al SSRN 1478214.
- 350 Raghavan, Manish, Barocas, Solon, Kleinberg, Jon e Levy, Karen. 2019. "Mitigating bias nello screening algoritmico dell'occupazione: valutazione di richieste e pratiche". arXiv prestampa *arXiv:1906.09208*.
- 351 Yao, Sirui e Huang, Bert. 2017. "Oltre la parità: obiettivi di equità per il filtraggio collaborativo". In Advances in Neural Information Processing Systems, pagine 2921–2930.
- 352 Mehrotra, Rishabh, Anderson, Ashton, Diaz, Fernando, Sharma, Amit, Wallach, Hanna e Yilmaz, Emine. 2017. "Auditing dei motori di ricerca per la soddisfazione differenziale tra i dati demografici". Nella Conferenza internazionale sul World Wide Web, pagine 626–633.
- 353 Susser, Daniel, Roessler, Beate, e Nissenbaum, Helen. 2018. "Manipolazione online: nascosta influenze in un mondo digitale". Disponibile al SSRN 3306006.
- 354 Bagwell, Kyle. 2007. "L'analisi economica della pubblicità". Manuale di organizzazione industriale, 3:1701–1844.
- 355 Coltrane, Scott e Messineo, Melinda. 2000. "La perpetuazione del sottile pregiudizio: Razza e immagini di genere nella pubblicità televisiva degli anni '90". Ruoli sessuali, 42 (5-6): 363–389.
- 356 Angwin, Julia, Varner, Madeleine e Tobin, Ariana. 2017. "Facebook ha consentito agli inserzionisti di raggiungere gli "odiatori degli ebrei"". ProPublica. <https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters> .
- 357 Lambrecht, Anja e Tucker, Caterina. 2019. "Distorsione algoritmica? uno studio empirico sull'apparente discriminazione basata sul genere nella visualizzazione di annunci di carriera staminali". Scienze gestionali.
- 358 Datta, Amit, Tschantz, Michael Carl e Datta, Anupam. 2015. "Esperimenti automatizzati sull'annuncio impostazioni sulla privacy". Proc. Tecnologie per il miglioramento della privacy (PET), 2015(1):92–112.
- 359 Andreou, Athanasios, Goga, Oana, Gummadi, Krishna, Loiseau, Patrick, e Mislove, Alan. 2017. "Adanalista". <https://adanalyst.mpi-sws.org/>.
- 360 Hutson, Jevan A, Taft, Jessie G, Barocas, Solon e Levy, Karen. 2018. "Debiasare il desiderio: affrontare pregiudizi e discriminazioni sulle piattaforme intime". Atti dell'ACM sull'interazione uomo- computer, 2 (CSCW):73.
- 361 Ayres, Ian, Banaji, Mahzarin e Jolls, Christine. 2015. "Effetti della corsa su eBay". Il RAND Journal of Economics, 46(4):891–917.
- 362 Lee, Min Kyung, Kusbit, Daniel, Metsky, Evan e Dabbish, Laura. 2015. "Lavorare con le macchine: l'impatto della gestione algoritmica e basata sui dati sui lavoratori umani". In Conference on Human Factors in Computing Systems (CHI), pagine 1603–1612. ACM.
- 363 Edelman, Benjamin, Luca, Michael e Svirsky, Dan. 2017. "La discriminazione razziale nell'economia della condivisione : prove da un esperimento sul campo". American Economic Journal: Economia applicata, 9(2):1–22.
- 364 Thebault-Spieker, Jacob, Terveen, Loren e Hecht, Brent. 2017. "Verso una comprensione geografica dell'economia della condivisione: pregiudizi sistemicci in uberx e taskrabbit". Transazioni ACM sull'interazione uomo-computer (TOCHI), 24(3):1–40.
- 365 Ge, Yanbo, Knittel, Christopher R, MacKenzie, Don, e Zoepf, Stephen. 2016. " Discriminazioni razziali e di genere nelle imprese delle reti di trasporto". Rapporto tecnico, Ufficio nazionale di ricerca economica.
- 366 Levy, Karen e Barocas, Solone. 2017. "Progettare contro la discriminazione nei mercati online". Berkeley Tech. LJ, 32:1183.
- 367 Tjaden, Jasper Dag, Schwemmer, Carsten e Khadjavi, Menusch. 2018. "Ride with me: discriminazione etnica, mercati sociali ed economia della condivisione". Rivista sociologica europea, 34(4):418–432 .
- 368 Muthukumar, Vidya, Pedapati, Tejaswini, Ratha, Nalini, Sattigeri, Prasanna, Wu, Chai-Wah, Kingsbury, Brian, Kumar, Abhishek, Thomas, Samuel, Mojsilovic, Aleksandra e Varshney, Kush R. 2018. "Understanding unequal accuratezza della classificazione di genere dalle immagini dei volti". arXiv prestampa *arXiv:1812.00099*.

- 369 Harvey, Adam e LaPlace, Jules. 2019. "Megapixel: origini, etica e implicazioni sulla privacy di set di dati di immagini di riconoscimento facciale pubblicamente disponibili".
- 370 Robertson, Ronald E, Jiang, Shan, Joseph, Kenneth, Friedland, Lissa, Lazer, David e Wilson, Christo. 2018. "Verifica dei pregiudizi del pubblico partigiano all'interno della ricerca Google". Atti dell'ACM sull'interazione uomo-computer, 2 (CSCW):148.
- 371 D'Onfro, Jillian. 2019. "Google testa le modifiche al suo algoritmo di ricerca; come funziona la ricerca". <https://www.cnbc.com/2018/09/17/google-tests-changes-to-its-search-algorithm-how-search-works.html>.
- 372 Hannak, Aniko, Sapiezynski, Piotr, Molavi Kakhki, Arash, Krishnamurthy, Balachander, Lazer, David, Mislove, Alan e Wilson, Christo. 2013. "Misurare la personalizzazione della ricerca web". In Conferenza internazionale sul World Wide Web, pagine 527–538. ACM.
- 373 Tripodi, Francesca. 2018. "Alla ricerca di fatti alternativi: analisi dell'inferenza scritturale nelle pratiche giornalistiche conservatrici". Dati e società, 29.
- 374 Golebiewski, M e Boyd, D. 2018. "Data voids: dove i dati mancanti possono essere facilmente sfruttati". Dati e società, 29.
- 375 Valentino-DeVries, Jennifer, Singer-Vine, Jeremy e Soltani, Ashkan. 2012. "I siti web variano prezzi, offerte in base alle informazioni degli utenti". Wall Street Journal, 10:60–68.
- 376 Ojala, Markus e Garriga, Gemma C. 2010. "Test di permutazione per studiare le prestazioni del classificatore ". Journal of Machine Learning Research, 11 (giugno): 1833–1863.
- 377 Venkatadri, Giridhari, Lucherini, Elena, Sapiezynski, Piotr, e Mislove, Alan. 2019. "Indagare le fonti dei pii utilizzati nella pubblicità mirata di Facebook". Atti sulle tecnologie per il miglioramento della privacy, 2019(1):227–244.
- 378 Bashir, Muhammad Ahmad, Arshad, Sajjad, Robertson, William, e Wilson, Christo. 2016. "Tracciamento dei flussi di informazioni tra gli scambi di annunci utilizzando annunci con retargeting". Nel Simposio sulla sicurezza USENIX 16, pagine 481–496.
- 379 Singer-Vine, Jeremy, Valentino-DeVries, Jennifer e Soltani, Ashkan. 2012. "Come la rivista ha testato prezzi e offerte online". Giornale di Wall Street. <http://blogs.wsj.com/digits/2012/12/23/how-the-journal-tested-prices-and-deals-online>.
- 380 Chen, Le, Mislove, Alan, e Wilson, Christo. 2015. "Sbirciare sotto il cofano di Uber". In Atti della conferenza sulla misurazione di Internet del 2015 , pagine 495–508. ACM.
- 381 Salganik, Matteo. 2019. A poco a poco: La ricerca sociale nell'era digitale. Stampa dell'Università di Princeton.
- 382 Bennett, James e Lanning, Stan. 2007. "Il premio netflix". Negli Atti della Coppa KDD e workshop, volume 2007, pagina 35. New York, NY, USA.
- 383 Chaney, Allison JB, Stewart, Brandon M e Engelhardt, Barbara E. 2018. "Come il confondimento algoritmico nei sistemi di raccomandazione aumenta l'omogeneità e diminuisce l'utilità". Nella conferenza ACM sui sistemi di raccomandazione, pagine 224–232. ACM.
- 384 Obermeyer, Ziad, Powers, Brian, Vogeli, Christine e Mullainathan, Sendhil. 2019. "Dissezione dei pregiudizi razziali in un algoritmo utilizzato per gestire la salute delle popolazioni". Scienza, 366 (6464): 447–453.
- 385 Chouldechova, Alexandra, Benavides-Prado, Diana, Fialko, Oleksandr e Vaithianathan, Rhema. 2018. "Un caso di studio sul processo decisionale assistito da algoritmi nelle decisioni di screening della hotline per il maltrattamento sui minori". In Conferenza su equità, responsabilità e trasparenza, pagine 134–148.
- 386 Narayanan, Arvind. 2022. "I limiti dell'approccio quantitativo alla discriminazione". Giacomo Baldwin Lecture [trascrizione], Princeton University.
- 387 Gaddis, S. Michele. 2018. "Un'introduzione agli studi di audit nelle scienze sociali". Negli studi di audit: Dietro le quinte con teoria, metodo e sfumature, pagine 3–44. Springer.
- 388 Vecchione, Briana, Levy, Karen, e Barcas, Solon. 2021. "Audit algoritmico e giustizia sociale: lezioni dalla storia degli studi di audit". In Equità e accesso in algoritmi, meccanismi e ottimizzazione, pagine 1–9.
- 389 Brundage, Miles, Avin, Shahar, Wang, Jasmine, Belfield, Haydn, Krueger, Gretchen, Hadfield, Gillian, Khlaaf, Heidy, Yang, Jingying, Toner, Helen, Fong, Ruth, et al.. 2020. "Verso affidabile ai Development: meccanismi per sostenere affermazioni verificabili". arXiv prestampa <arXiv:2004.07213>.
- 390 Wu, Tim. 2010. L'interruttore principale: l'ascesa e la caduta degli imperi dell'informazione. Annata.
- 391 Gillespie, Tarleton. 2010. "La politica delle 'piattaforme'". Nuovi media e società, 12(3):347–364.
- 392 ___. 2018. Custodi di Internet: piattaforme, moderazione dei contenuti e decisioni nascoste che modellano i social media. Stampa dell'Università di Yale.

- 393 Klonick, Kate. 2017. "I nuovi governatori: le persone, le regole e i processi che governano online discorso". *Harv. L. Rev.*, 131:1598.
- 394 Cook, Cody, Diamond, Rebecca, Hall, Jonathan, List, John A e Oyer, Paul. 2018. "Il divario di genere nei guadagni nella gig economy: prove di oltre un milione di conducenti di rideshare". Rapporto tecnico , Ufficio nazionale di ricerca economica.
- 395 Vincitore, Langdon. 2017. Gli artefatti hanno politica? Routledge.
- 396 Bullard, Robert Doyle, Johnson, Glenn Steve e Torres, Angel O. 2004. Rapina in autostrada: Razzismo nei trasporti e nuove vie verso l'equità. Stampa dell'estremità sud.
- 397 Piccolo, Mario L e Cercapersone, Devah. 2020. "Prospettive sociologiche sulla discriminazione razziale". Giornale delle prospettive economiche, 34(2):49–67.
- 398 Altman, Andrea. 2020. "Discriminazione". In Edward N. Zalta (a cura di), *The Stanford Encyclopedia of Filosofia*. Laboratorio di ricerca sulla metafisica, Università di Stanford, edizione inverno 2020.
- 399 Giglio Hu. 2020. "Effetti diretti". <https://phenomenalworld.org/analysis/direct-effects>.
- 400 contributori di Wikipedia. 2021. "Matrimonio tra persone dello stesso sesso – Wikipedia, l'enciclopedia libera".
- 401 Rothstein, Richard. 2017. Il colore della legge: una storia dimenticata di come il nostro governo ha segregato l'America. Pubblicazione di Liveright.
- 402 Fellner, Jamie. 2009. "Razza, droga e applicazione della legge negli Stati Uniti". Stan. L. e Pol'y Rev., 20:257.
- 403 Mou, Ted. 2002. "I lavoratori neri perdono la connessione? l'effetto della distanza spaziale e segnalazioni di dipendenti sulla segregazione razziale tra aziende". *Demografia*, 39 (3): 507–528.
- 404 Leslie, Sarah-Jane, Cimpian, Andrei, Meyer, Meredith e Freeland, Edward. 2015. "Le aspettative di brillantezza sono alla base delle distribuzioni di genere nelle discipline accademiche". *Scienza*, 347 (6219): 262–265.
- 405 Bian, Lin, Leslie, Sarah-Jane e Cimpian, Andrei. 2017. "Gli stereotipi di genere sulle capacità intellettuali emergono presto e influenzano gli interessi dei bambini". *Scienza*, 355 (6323): 389–391.
- 406 West, Candace e Zimmerman, Don H. 1987. "Doing gender". *Genere e società*, 1 (2): 125–151.
- 407 Valencia Caicedo, Felipe. 2019. "La missione: trasmissione del capitale umano, persistenza economica, e cultura in Sud America". Il giornale trimestrale di economia, 134(1):507–556.
- 408 Dell, Melissa. 2010. "Gli effetti persistenti della mita mineraria del Perù". *Econometrica*, 78(6):1863–1903.
- 409 Barone, Guglielmo e Mocetti, Sauro. 2016. "La mobilità intergenerazionale nel lunghissimo periodo: Firenze 1427-2011". Banca d'Italia Temi di Discussione n. 1060.
- 410 Davidai, Shai e Gilovich, Thomas. 2015. "Costruire un'America più mobile: un quintile di reddito alla volta". Prospettive sulla scienza psicologica, 10(1):60–71.
- 411 Chetty, Raj, Hendren, Nathaniel, Jones, Maggie R e Porter, Sonya R. 2020. "Razza e opportunità economiche negli Stati Uniti: una prospettiva intergenerazionale". Il giornale trimestrale di economia, 135(2):711–783.
- 412 Kochhar, Rakesh e Cilluffo, Anthony. 2018. "Risultati chiave sull'aumento della disegualanza di reddito all'interno dei gruppi razziali ed etnici americani". Centro di ricerca Pew.
- 413 Derenoncourt, Ellora, Kim, Chi Hyun, Kuhn, Moritz, e Schularick, Moritz. 2022. "La ricchezza di due nazioni: il divario di ricchezza razziale negli Stati Uniti, 1860-2020". Rapporto tecnico, Ufficio nazionale di ricerca economica.
- 414 Kraus, Michael W, Onyeador, Ivuoma N, Daumeyer, Natalie M, Rucker, Julian M e Richeson, Jennifer A. 2019. "The misperception of racial economic inequality". Prospettive sulla scienza psicologica, 14(6):899–921.
- 415 Semega, Jessica L, Fontenot, Kayla R e Kollar, Melissa A. 2017. "Reddito e povertà nei Stati Uniti: 2016". Rapporti sulla popolazione attuale, (P60-259).
- 416 Pendall, Rolf e Hedman, Carl. 2015. "Mondi a parte: disegualanza tra i quartieri più e meno ricchi d'America ". Istituto Urbano.
- 417 Merrill, Jeremy. 2020. "Facebook vende ancora annunci discriminatori?"
- 418 Gandy, Oscar H. 2016. Venire a patti con il caso: Engaging razionale discriminazione e cumulativa svantaggio. Routledge.
- 419 Hellman, Debora. 2020. "Sesso, causalità e algoritmi: come vieta la parità di protezione aggravando l'ingiustizia precedente". Revisione giuridica dell'Università di Washington, 98(2):481–523.
- 420 Myrdal, Gunnar. 2017. Un dilemma americano: il problema dei negri e la democrazia moderna, volume 2. Routledge.

- 421 Boyd, D. 2012. "Il volo dei bianchi nei pubblici in rete: come la razza e la classe hanno modellato il coinvolgimento degli adolescenti americani con MySpace e Facebook. nakamura I, chow-white pa, eds. corsa dopo internet". Corsa dopo Internet, pagine 203–222.
- 422 Kahneman, Daniel, Rosenfield, AM, Gandhi, L, e Blaser, T. 2016. "Rumore: come superare il costo elevato e nascosto di un processo decisionale incoerentehttps". Revisione aziendale di Harvard.
- 423 Kleinberg, Jon e Raghavan, Manish. 2021. "Monocultura algoritmica e welfare sociale". Atti dell'Accademia Nazionale delle Scienze, 118(22):e2018340118.
- 424 O'Neil, Cathy. 2016. "Come gli algoritmi governano la nostra vita lavorativa". Il Guardiano, 16.
- 425 Gillespie, Tarleton. 2020. "Moderazione dei contenuti, intelligenza artificiale e questione di scala". Big Data e Società, 7(2):2053951720943234.
- 426 Kalluri, Ria. 2019. "I valori del machine learning". Workshop NeurIPS Queer nell'intelligenza artificiale.
- 427 Lokugamage, A, Taylor, S e Rayner, C. 2020. "Le esperienze dei pazienti con "longcovid" mancano nella narrativa del servizio sanitario nazionale". BMJ.
- 428 Callard, Felicity e Perego, Elisa. 2021. "Come e perché i pazienti hanno reso il covid lungo". Scienze sociali e medicina, 268:113426.
- 429 Sambasivan, Nithya e Veeraraghavan, Rajesh. 2022. "La dequalificazione delle competenze di dominio nello sviluppo dell'intelligenza artificiale". Nella conferenza CHI sui fattori umani nei sistemi informatici, pagine 1–14.
- 430 Lipsky, Michael. 2010. Burocrazia di strada: dilemmi dell'individuo nel servizio pubblico. Russell Fondazione Salvia.
- 431 . 2021. "Informatori: bug del software che mantiene centinaia di detenuti nelle carceri dell'Arizona oltre le date di rilascio". KJZZ.
- 432 . 2015. "Prigionieri statunitensi rilasciati anticipatamente a causa di un bug del software". Notizie della BBC.
- 433 Mulligan, Deirdre K e Bamberger, Kenneth A. 2019. "Appalti come politica: amministrazione processo per l'apprendimento automatico". Berkeley Tech. LJ, 34:773.
- 434 Barabas, Chelsea, Virza, Madars, Dinakar, Karthik, Ito, Joichi e Zittrain, Jonathan. 2018. "Interventi sulle previsioni: riformulare il dibattito etico per la valutazione del rischio attuariale". In Conferenza su equità, responsabilità e trasparenza, pagine 62–76. PMLR.
- 435 Akbar, Amna. 2020. "Un orizzonte abolizionista per la polizia (riforma)". California Law Review, 108(6).
- 436 Roberts, Dorothy. 2022. Lacerato: come il sistema di welfare infantile distrugge le famiglie nere e come l'abolizione può costruire un mondo più sicuro. Libri di base.
- 437 Karabel, Girolamo. 2005. Gli eletti: la storia nascosta dell'ammissione e dell'esclusione ad Harvard, Yale e Princeton. Houghton Mifflin Harcourt.
- 438 Harwell, Drew. 2019. "L'azienda di telecamere per campanelli ha collaborato con 400 forze di polizia, estendendo le preoccupazioni sulla sorveglianza". Washington Post.
- 439 Whittaker, Meredith, Crawford, Kate, Dobbe, Roel, Fried, Genevieve, Kaziunas, Elizabeth, Mathur, Varoon, West, Sarah Myers, Richardson, Rashida, Schultz, Jason e Schwartz, Oscar. 2018. Rapporto AI now 2018. AI Now Institute presso la New York University di New York.
- 440 Baroca, Solone e Levy, Karen. 2020. "Dipendenze dalla privacy". Washington L. Rev., 95:555.
- 441 Sloane, Mona, Moss, Emanuel, Awomolo, Olaitan e Forlano, Laura. 2020. "La partecipazione no una soluzione progettuale per l'apprendimento automatico". arXiv prestampa arXiv:2007.02423.
- 442 Hoffmann, Anna Lauren. 2019. "Dove l'equità fallisce: dati, algoritmi e limiti dell'antidiscorso sulla criminalità". Informazione, comunicazione e società, 22(7):900–915.
- 443 Jobin, Anna, lenca, Marcello, e Vayena, Effy. 2019. "Il panorama globale delle linee guida etiche dell'AI". Natura Macchina Intelligenza, 1(9):389–399.
- 444 Greene, Daniel, Hoffmann, Anna Lauren e Stark, Luke. 2019. "Migliore, più bello, più chiaro, più giusto: una valutazione critica del movimento per l'intelligenza artificiale etica e l'apprendimento automatico". In Atti della 52a conferenza internazionale delle Hawaii sulle scienze dei sistemi.
- 445 Judd, Sarah. 2020. "Attività per costruire la comprensione: come ai4all insegna l'intelligenza artificiale a diversi studenti delle scuole superiori". In Atti del 51° Simposio tecnico ACM sull'educazione informatica, pagine 633–634. 446 . 2021. "Annuncio del progetto sulla disparità dei lavoratori a contratto". Collaborazione azionaria tecnologica.
- 447 Cowgill, Bo, Dell'Acqua, Fabrizio, Deng, Samuel, Hsu, Daniel, Verma, Nakul e Chaintreau, Augustin. 2020. "Programmatori di parte? o dati distorti? un esperimento sul campo per rendere operativa l' etica dell'IA". In Conference on Economics and Computation, pagine 679–681.

- 448 Fiesler, Casey, Garrett, Natalie e Beard, Nathan. 2020. "Cosa insegniamo quando insegniamo l'etica tecnologica? un'analisi dei programmi". In Atti del 51° Simposio tecnico ACM sull'educazione informatica , pagine 289–295.
- 449 Martin, C Dianne, Huff, Chuck, Gotterbarn, Donald e Miller, Keith. 1996. "Implementare a decimo filone del curriculum CS". Comunicazioni dell'ACM, 39(12):75–84.
- 450 contributori di Wikipedia. 2021. "Professionista certificato nello sviluppo di software: Wikipedia, the enciclopedia libera".
- 451 Birhane, Abeba, Kalluri, Pratyusha, Card, Dallas, Agnew, William, Dotan, Ravit e Bao, Michelle. 2022. "I valori codificati nella ricerca sul machine learning". In Conferenza su equità, responsabilità e trasparenza, pagine 173–184.
- 452 Nanayakkara, Priyanka, Hullman, Jessica, e Diakopoulos, Nicholas. 2021. "Disimballare le conseguenze espresse della ricerca sull'intelligenza artificiale in dichiarazioni di impatto più ampie". arXiv prestampa *arXiv:2105.04760*.
- 453 Kasy, Massimiliano e Abebe, Rediet. 2021. "Equità, uguaglianza e potere nel processo decisionale algoritmico ". In Conferenza su equità, responsabilità e trasparenza, pagine 576–586.
- 454 Paluck, Elizabeth Levy e Green, Donald P. 2009. "Riduzione del pregiudizio: cosa funziona? una revisione e una valutazione della ricerca e della pratica". Revisione annuale della psicologia, 60: 339–367.
- 455 Paluck, Elizabeth Levy, Porat, Roni, Clark, Chelsey S e Green, Donald P. 2020. "Pregiudizio riduzione: progresso e sfide". Revisione annuale di psicologia, 72.
- 456 Chohlas-Wood, Alex, Nudell, Joe, Lin, Zhiyuan Jerry, Nyarko, Julian e Goel, Sharad. 2020. "Giustizia cieca: mascherare algoritmicamente la razza nelle decisioni di imputazione". Relazione tecnica, Relazione tecnica.
- ⁴⁵⁷ . 2021. "Indagini pre-assunzione e stato civile o numero di figli". Commissione statunitense per le pari opportunità di lavoro.
- 458 Squires, Gregory D et al.. 1994. Capitale e comunità in bianco e nero: le intersezioni di razza, classe e sviluppo ineguale. Suny Press.
- 459 Raso, Jennifer. 2017. "Lo spostamento come regolamentazione: nuove tecnologie di regolamentazione e processo decisionale in prima linea nelle opere dell'Ontario". Giornale canadese di diritto e società, 32(1):75–95.
- 460 Brayne, Sarah. 2020. Prevedere e sorvegliare: dati, discrezione e futuro della polizia. Università di Oxford Stampa, Stati Uniti.
- 461 Frankel, Marvin E. 1973. "Sentenze penali: legge senza ordine".
- 462 Palamar, Joseph J, Davies, Shelby, Ompad, Danielle C, Cleland, Charles M e Weitzman, Michael. 2015. "Cocaina in polvere e uso di crack negli Stati Uniti: un esame del rischio di arresto e delle disparità socioeconomiche nell'uso". Dipendenza da droga e alcol, 149: 108–116.
- 463 Google re:gruppo di lavoro. 2021. "Guida: assunzione da parte del comitato". <https://rework.withgoogle.com/stampa/guide/6053596147744768/>.
- 464 Natasha Tiku. 2021. "L'approccio di Google alle scuole storicamente nere aiuta a spiegare perché esistono pochi ingegneri neri nella grande tecnologia". Washington Post.
- 465 Edelman, Lauren B. 2005. "Il diritto al lavoro: la costruzione endogena dei diritti civili". Nel Manuale sulla ricerca sulla discriminazione sul lavoro, pagine 337–352. Springer.
- 466 Soper, S. 2021. "Licenziato dal bot su Amazon:'sei tu contro la macchina.'". Bloomberg.
- 467 Dynarski, Susan, Libassi, CJ, Michelmore, Katherine e Owen, Stephanie. 2018. "Colmare il divario: l'effetto di una promessa mirata e senza tasse scolastiche sulle scelte universitarie degli studenti ad alto rendimento e a basso reddito". Rapporto tecnico, Ufficio nazionale di ricerca economica.
- 468 Antecol, Heather, Bedard, Kelly e Stearns, Jenna. 2018. "Uguali ma ingiusti: chi trae vantaggio dalle politiche di sospensione dell'orologio di ruolo neutrali rispetto al genere?" American Economic Review, 108(9):2420–41.
- 469 Fishbane, Alissa, Ouss, Aurelie e Shah, Anuj K. 2020. "I nudge comportamentali riducono il fallimento comparire in tribunale". Scienza, 370(6517).
- 470 Hellman, Debora. 2016. "Due concetti di discriminazione". Va. L. Rev., 102:895.
- 471 Stevenson, Megan T e Mayson, Sandra G. 2021. "Detenzione preventiva e valore della libertà". Documento di ricerca sul diritto pubblico e sulla teoria giuridica della Virginia, (2021-14).
- 472 Eaglin, Jessica M. 2017. "Costruire il rischio di recidiva". Emory LJ, 67:59.
- 473 Verde, Ben. 2021. "La scienza dei dati come azione politica: radicare la scienza dei dati in una politica di giustizia". Giornale di informatica sociale, 2(3):249–265.
- 474 Kohler-Hausmann, Issa. 2018. "Misdemeanorland". Nel paese dei misfatti. Università di Princeton Premere.

- 475 Paluck, Elizabeth Levy, Green, Seth A e Green, Donald P. 2019. "L'ipotesi di contatto rivalutato". *Politica pubblica comportamentale*, 3(2):129–158.
- 476 Lundberg, Shelly e Startz, Richard. 1998. "Sulla persistenza della disuguaglianza razziale". *Giornale di economia del lavoro*, 16(2):292–323.
- 477 Bowles, Samuel e Sethi, Rajiv. 2006. "Segregazione sociale e dinamica della disuguaglianza di gruppo".
- 478 Loury, Glenn C. 1976. "Una teoria dinamica delle differenze di reddito razziale". *Relazione tecnica*, Documento di discussione.
- 479 Massey, Douglas S, Rothwell, Jonathan, e Domina, Thurston. 2009. "Le mutevoli basi della segregazione negli Stati Uniti". *Gli Annali dell'Accademia Americana di Scienze Politiche e Sociali*, 626(1):74–90.
- 480 Li, Xiaochang e Mills, Mara. 2019. "Caratteristiche vocali: dall'identificazione vocale al riconoscimento vocale" zione a macchina". *Tecnologia e cultura*, 60(2):S129–S160.
- 481 Libermann, Marco. 2010. "Fred Jelinek". *Linguistica computazionale*, 36(4):595–599.
- 482 Chiesa, Kenneth Ward. 2018. "Tendenze emergenti: un omaggio a Charles Wayne". *Linguaggio naturale Ingegneria*, 24(1):155–160.
- 483 Liberman, Mark e Wayne, Charles. 2020. "Tecnologia del linguaggio umano". *Rivista AI*, 41(2).
- 484 Garofolo, John S, Lamel, Lori F, Fisher, William M, Fiscus, Jonathan G e Pallett, David S. 1993. "Darpa timit corpus cd-rom del discorso continuo acustico-fonetico. disco parlato 1-1.1". Rapporto tecnico NASA STI/Recon n. 93:27403.
- 485 Koenecke, Allison, Nam, Andrew, Lake, Emily, Nudell, Joe, Quartey, Minnie, Mengesha, Zion, Toups, Connor, Rickford, John R, Jurafsky, Dan e Goel, Sharad. 2020. "Disparità razziali nel riconoscimento vocale automatizzato". Atti dell'Accademia Nazionale delle Scienze, 117(14):7684–7689.
- 486 Langley, Pat. 2011. "La scienza in evoluzione dell'apprendimento automatico".
- 487 LeCun, Yann, Bottou, Léon, Bengio, Yoshua, e Haffner, Patrick. 1998. "Apprendimento basato sui gradienti applicata al riconoscimento dei documenti". Atti dell'IEEE, 86(11):2278–2324.
- 488 Grother, Patrick J. 1995. "Database speciale Nist 19". Database di moduli e caratteri stampati a mano, Istituto nazionale di standard e tecnologia, pagina 10.
- 489 Yadav, Chhavi e Bottou, Léon. 2019. "Caso freddo: le cifre ministe perdute". arXiv prestampa arXiv:1905.10498.
- 490 DeCoste, Dennis e Schölkopf, Bernhard. 2002. "Addestramento di macchine vettoriali a supporto invariante". Apprendimento automatico, 46(1):161–190.
- 491 Bromley, J e Sackinger, E. 1991. "Classificatori di reti neurali e k-vicino più vicino". *Tecnica del rapporto*, pagine 11359–910819.
- 492 Miller, George A. 1998. WordNet: un database lessicale elettronico. Stampa del MIT.
- 493 Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai e Fei-Fei, Li. 2009. "Imagenet: un database di immagini gerarchico su larga scala". Nel 2009 conferenza IEEE sulla visione artificiale e il riconoscimento di modelli, pagine 248–255. ieee.
- 494 Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al.. 2015. "Imagenet large sfida del riconoscimento visivo su larga scala". *Rivista internazionale di visione artificiale*, 115(3):211–252.
- 495 Gray, Mary L e Suri, Siddharth. 2019. Lavoro fantasma: come impedire alla Silicon Valley di costruirne una nuova sottoclasse globale. *Libri di Eamon Dolan*.
- 496 Narayanan, Arvind e Shmatikov, Vitaly. 2008. "Robusta de-anonimizzazione di grandi set di dati sparsi". Nel 2008 IEEE Symposium on Security and Privacy (sp 2008), pagine 111–125. IEEE.
- 497 Dwork, Cynthia, Smith, Adam, Steinke, Thomas e Ullman, Jonathan. 2017. "Esposto! un sondaggio di attacchi ai dati privati". *Revisione annuale delle statistiche e della sua applicazione*, 4: 61–84.
- 498 Dwork, Cynthia, Roth, Aaron, et al.. 2014. "I fondamenti algoritmici della privacy differenziale". Fondamenti e tendenze nell'informatica teorica, 9 (3-4): 211–407.
- 499 Fisher, Ronald A. 1936. "L'uso di misurazioni multiple in problemi tassonomici". *Annali di eugenetica*, 7(2):179–188.
- 500 Evans, Richard. 2020. "Ra Fisher e la scienza dell'odio".
- 501 Louça, Francisco. 2009. "Emancipazione attraverso l'interazione: come l'eugenetica e la statistica convergevano e divergevano". *Giornale di storia della biologia*, 42(4):649–684.
- 502 Mitchell, Tom M. 1980. La necessità di pregiudizi nell'apprendimento delle generalizzazioni. Dipartimento di Informatica Scienza, Laboratorio di ricerca informatica. . .

- 503 Breiman, Leo et al.. 2001. "Modellazione statistica: le due culture (con commenti e contoreplica dell'autore)". *Scienza statistica*, 16(3):199–231.
- 504 Rosenblatt, Frank. 1960. "Esperimenti di simulazione del percettrone". *Atti dell'IRE*, 48(3):301–309.
- 505 Langley, Pat. 1988. "L'apprendimento automatico come scienza sperimentale".
- 506 Funk, Simon. 2006. "Provalo a casa". <http://sifter.org/~simon/journal/2006>.
- 507 Billsus, Daniel, Pazzani, Michael J, et al.. 1998. "Apprendimento dei filtri di informazione collaborativi". In *Icmi*, volume 98, pagine 46–54.
- 508 Deerwester, Scott, Dumais, Susan T, Furnas, George W, Landauer, Thomas K, e Harshman, Richard. 1990. "Indicizzazione mediante analisi semantica latente". *Giornale della società americana per la scienza dell'informazione*, 41(6):391–407.
- 509 Koren, Yehuda, Bell, Robert, e Volinsky, Chris. 2009. "Tecniche di fattorizzazione di matrici per sistemi di raccomandazione". *Computer*, 42(8):30–37.
- 510 Recht, Benjamin, Roelofs, Rebecca, Schmidt, Ludwig e Shankar, Vaishaal. 2019. "I classificatori Imagenet si generalizzano a Imagenet?" Nella conferenza internazionale sull'apprendimento automatico, pagine 5389–5400. PMLR.
- 511 Miller, John P, Taori, Rohan, Raghunathan, Aditi, Sagawa, Shiori, Koh, Pang Wei, Shankar, Vaishaal, Liang, Percy, Carmon, Yair e Schmidt, Ludwig. 2021. "Precisione sulla linea: sulla forte correlazione tra generalizzazione fuori distribuzione e generalizzazione in distribuzione". Nella conferenza internazionale sull'apprendimento automatico, pagine 7721–7735. PMLR.
- 512 Cortes, Corinna e Vapnik, Vladimir. 1995. "Reti di vettori di supporto". *Apprendimento automatico*, 20(3):273–297.
- 513 Henrich, Joseph, Heine, Steven J e Norenzayan, Ara. 2010. "Le persone più strane del mondo?" *Scienze comportamentali e del cervello*, 33 (2-3): 61–83.
- 514 Crawford, Kate e Paglen, Trevor. 2019. "Excavating ai: la politica dei training set per il machine learning". Scavo AI (www.excavating.ai).
- 515 Gao, Leo, Biderman, Stella, Black, Sid, Golding, Laurence, Hoppe, Travis, Foster, Charles, Phang, Jason, He, Horace, Thite, Anish, Nabeshima, Noa, et al.. 2020. "The pile : Un set di dati da 800 GB di testi diversi per la modellazione del linguaggio". arXiv prestampa [arXiv:2101.00027](https://arxiv.org/abs/2101.00027).
- 516 Veale, Michael, Binns, Ruben e Edwards, Lilian. 2018. "Algoritmi che ricordano: attacchi con inversione di modello e normativa sulla protezione dei dati". *Transazioni filosofiche della Royal Society A: Scienze matematiche, fisiche e ingegneristiche*, 376(2133):20180083.
- 517 Duda, Richard O, Hart, Peter E e Stork, David G. 1973. *Classificazione dei modelli e analisi delle scene*, volume 3. Wiley New York.
- 518 Hastie, Trevor, Tibshirani, Robert e Friedman, Jerome. 2017. *Gli elementi dell'apprendimento statistico: data mining, inferenza e previsione*. Springer.
- 519 Blum, Avrim e Hardt, Moritz. 2015. "La scala: una classifica affidabile per le competizioni di machine learning". Nella conferenza internazionale sull'apprendimento automatico, pagine 1006–1014. PMLR.
- 520 Dehghani, Mostafa, Tay, Yi, Gritsenko, Alexey A, Zhao, Zhe, Housby, Neil, Diaz, Fernando, Metzler, Donald e Vinyals, Oriol. 2021. "La lotteria di riferimento". arXiv prestampa [arXiv:2107.07002](https://arxiv.org/abs/2107.07002).
- 521 Koch, Bernard, Denton, Emily, Hanna, Alex e Foster, Jacob G. 2021. "Ridotto, riutilizzato e riciclato: la vita di un set di dati nella ricerca sull'apprendimento automatico". arXiv prestampa [arXiv:2112.01716](https://arxiv.org/abs/2112.01716).
- 522 Koh, Pang Wei, Sagawa, Shiori, Marklund, Henrik, Xie, Sang Michael, Zhang, Marvin, Balsubra-mani, Akshay, Hu, Weihua, Yasunaga, Michihiro, Phillips, Richard Lanas, Gao, Irena, et al.. 2020. "Wilds: un punto di riferimento dei cambiamenti di distribuzione in natura". arXiv prestampa [arXiv:2012.07421](https://arxiv.org/abs/2012.07421).
- 523 Branwen, Gwern. 2011. "La leggenda metropolitana del carro armato della rete neurale".
- 524 Kaufman, Shachar, Rosset, Saharon, Perllich, Claudia e Stitelman, Ori. 2012. "Perdita nel data mining: formulazione, rilevamento ed evitamento". *Transazioni ACM sulla scoperta della conoscenza dai dati (TKDD)*, 6(4):1–21.
- 525 Marie, Benjamin, Fujita, Atsushi, e Rubino, Raffaello. 2021. "Credibilità scientifica della ricerca sulla traduzione automatica: una meta-valutazione di 769 articoli". arXiv prestampa [arXiv:2106.15195](https://arxiv.org/abs/2106.15195).
- 526 Bouthillier, Xavier, Delaunay, Pierre, Bronzi, Mirko, Trofimov, Assya, Nichyporuk, Brennan, Szeto, Justin, Mohammadi Sepahvand, Nazanin, Raff, Edward, Madan, Kanika, Voleti, Vikram, et al.. 2021. "Contabilità della varianza nei benchmark di apprendimento automatico". *Atti di machine learning e sistemi*, 3.

- 527 Saitta, Lorenza e Neri, Filippo. 1998. "L'apprendimento nel "mondo reale"". *Apprendimento automatico*, 30(2):133–163.
- 528 Salzberg, Steven L. 1999. "Sul confronto dei classificatori: una critica della ricerca e dei metodi attuali". *Estrazione dei dati e scoperta della conoscenza*, 1(1):1–12.
- 529 Radin, Joanna. 2017. ""nativi digitali": come le storie mediche e indigene contano per i grandi dati". *Osiride*, 32(1):43–64.
- 530 Kapoor, Sayash e Narayanan, Arvind. 2022. "Leakage e crisi di riproducibilità nella scienza basata sul ml". arXiv prestampa *arXiv:2207.07048*.
- 531 Prabhu, Vinay Uday e Birhane, Abeba. 2020. "Set di dati di immagini di grandi dimensioni: una vittoria di Pirro per il computer visione?" arXiv prestampa *arXiv:2006.16923*.
- 532 Yang, Kaiyu, Qinami, Klint, Fei-Fei, Li, Deng, Jia e Russakovsky, Olga. 2020. "Verso set di dati più equi: filtraggio e bilanciamento della distribuzione del sottoalbero delle persone nella gerarchia ImageNet". In Conferenza su equità, responsabilità e trasparenza, pagine 547–558.
- 533 Bender, Emily M, Gebru, Timnit, McMillan-Major, Angelina e Shmargaret. 2021. "Sui pericoli dei pappagalli stocastici: i modelli linguistici possono essere troppo grandi?" In Conferenza su equità, responsabilità e trasparenza, pagine 610–623.
- 534 Paullada, Amandalynne, Raji, Inioluwa Deborah, Bender, Emily M, Denton, Emily e Hanna, Alex. 2020. "Dati e i suoi (dis)contenuti: un'indagine sullo sviluppo e l'uso di set di dati nella ricerca sull'apprendimento automatico". arXiv prestampa *arXiv:2012.05345*.
- 535 Gebru, Timnit, Morgenstern, Jamie, Vecchione, Briana, Vaughan, Jennifer Wortman, Wallach, Hanna, Daumé III, Hal, e Crawford, Kate. 2018. "Schede tecniche per dataset". arXiv prestampa *arXiv:1803.09010*.
- 536 Jo, Eun Seo e Gebru, Timnit. 2020. "Lezioni dagli archivi: strategie per la raccolta di dati socioculturali nell'apprendimento automatico". In Conferenza su equità, responsabilità e trasparenza, pagine 306–316.
- 537 Wang, Angelina, Narayanan, Arvind e Russakovsky, Olga. 2020. "Revise: uno strumento per misurare e mitigare i bias nei set di dati visivi". Nella Conferenza europea sulla visione artificiale, pagine 733–751. Springer.
- 538 Gonen, Hila e Goldberg, Yoav. 2019. "Rossetto su un maiale: i metodi di debiasing nascondono sistematici pregiudizi di genere negli incorporamenti di parole ma non li rimuovono". arXiv prestampa *arXiv:1903.03862*.
- 539 Kuczmarski, James. 2018. "Ridurre i pregiudizi di genere in Google Translate". Blog di Google, 6.
- 540 Johnson, Melvin. 2020. "Un approccio scalabile per ridurre i pregiudizi di genere in Google Translate". Google Blog.
- 541 Destriero, Ryan e Caliskan, Aylin. 2021. "Le rappresentazioni di immagini apprese con un pre- addestramento non supervisionato contengono pregiudizi di tipo umano". In Conferenza su equità, responsabilità e trasparenza, pagine 701–713.
- 542 Huang, Gary B., Ramesh, Manu, Berg, Tamara e Learned-Miller, Erik. 2007. "Facce etichettate in natura: un database per studiare il riconoscimento facciale in ambienti non vincolati". Rapporto tecnico 07-49, Università del Massachusetts, Amherst.
- 543 Kumar, Neeraj, Berg, Alexander, Belhumeur, Peter N e Nayar, Shree. 2011. "Attributi visivi descrivibili per la verifica del volto e la ricerca di immagini". *Transazioni IEEE su Pattern Analysis e Machine Intelligence*, 33 (10): 1962–1977.
- 544 Jacobs, Abigail Z e Wallach, Hanna. 2021. "Misurazione ed equità". Nella Conferenza del Equità, responsabilità e trasparenza, pagine 375–385.
- 545 Messick, Samuele. 1998. "Validità del test: una questione di conseguenze". *Ricerca sugli indicatori sociali*, 45(1):35–44.
- 546 Dawes, Robyn M, Faust, David e Meehl, Paul E. 1989. "Giudizio clinico rispetto a quello attuariale". *Scienza*, 243 (4899): 1668–1674.
- 547 Chaaban, Ibrahim e Scheessele, Michael R. 2007. "Prestazioni umane nel database USPS". Rapporto, South Bend dell'Università dell'Indiana.
- 548 Lui, Kaiming, Zhang, Xiangyu, Ren, Shaoqing e Sun, Jian. 2015. "Approfondire i raddrizzatori: superare le prestazioni a livello umano sulla classificazione di Imagenet". Nella Conferenza internazionale sulla visione artificiale, pagine 1026–1034.
- 549 Shankar, Vaishaal, Roelofs, Rebecca, Mania, Horia, Fang, Alex, Recht, Benjamin, e Schmidt, Ludwig. 2020. "Valutazione dell'accuratezza della macchina su imangenet". Nella conferenza internazionale sull'apprendimento automatico , pagine 8634–8644. PMLR.

- 550 Herlocker, Jonathan L, Konstan, Joseph A, Terveen, Loren G e Riedl, John T. 2004. "Evaluating collaborative filtering Recommendations". *Transazioni ACM sui sistemi informativi (TOIS)*, 22(1):5–53.
- 551 Amatriain, Xavier e Basilico, Justin. 2012. "Consigli Netflix: Oltre le 5 stelle (parte 1)". Blog tecnico di Netflix, 6.
- 552 Blodgett, Su Lin, Lopez, Gilsinia, Olteanu, Alexandra, Sim, Robert, e Wallach, Hanna. 2021. "Stereotipizzazione del salmone norvegese: un inventario delle insidie nei set di dati di riferimento sull'equità". In Atti del 59° incontro annuale dell'Associazione per la linguistica computazionale e dell'11a conferenza congiunta internazionale sull'elaborazione del linguaggio naturale (volume 1: documenti lunghi), pagine 1004–1015 .
- 553 Bao, Michelle, Zhou, Angela, Zottola, Samantha, Brubach, Brian, Desmarais, Sarah, Horowitz, Aaron, Lum, Kristian e Venkatasubramanian, Suresh. 2021. "È complicato: la relazione disordinata tra set di dati rai e benchmark di equità algoritmica". arXiv prestampa *arXiv:2106.05498*.
- 554 Reichman, Nancy E, Teitler, Julien O, Garfinkel, Irwin e McLanahan, Sara S. 2001. "Famiglie fragili: campione e progettazione". *Revisione dei servizi per bambini e giovani*, 23 (4-5): 303–326.
- 555 Dwork, Cynthia, Feldman, Vitaly, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer e Roth, Aaron. 2015. "Il ostacolo riutilizzabile: preservare la validità nell'analisi adattiva dei dati". *Scienza*, 349 (6248): 636–638.
- 556 Boyd, Danah e Crawford, Kate. 2012. "Domande critiche per i big data: provocazioni per un fenomeno culturale, tecnologico e accademico". *Informazione, comunicazione e società*, 15(5):662–679.
- 557 Tufekci, Zeynep. 2014. "Grandi domande per i big data dei social media: rappresentatività, validità e altre insidie metodologiche". In Conferenza su Web e Social Media, volume 8. 558 —. 2014. "Ingegneria del pubblico: Big Data, sorveglianza e politica computazionale". Primo lunedì.
- 559 Couldry, Nick e Mejias, Ulises A. 2019. "Colonialismo dei dati: ripensare la relazione dei big data con il soggetto contemporaneo". *Televisione e nuovi media*, 20(4):336–349.
- 560 Olteanu, Alexandra, Castillo, Carlos, Diaz, Fernando e Kýcýman, Emre. 2019. "Dati sociali: pregiudizi, insidie metodologiche e confini etici". *Frontiere nei Big Data*, 2:13.
- 561 Crawford, Kate. 2021. L'Atlante dell'intelligenza artificiale. Stampa dell'Università di Yale.
- 562 Liberman, Marc. 2015. "La ricerca riproducibile e il metodo del compito comune". Conferenza della Fondazione Simmons <https://www.fondazione simon. org/lecture/reproducible-research-and-the-common-task-method>, 2.
- 563 Paullada, Amandalynne, Raji, Inioluwa Deborah, Bender, Emily M, Denton, Emily e Hanna, Alex. 2021. "Dati e i suoi (dis)contenuti: un'indagine sullo sviluppo e l'uso di set di dati nella ricerca sull'apprendimento automatico". *Modelli*, 2(11):100336.
- 564 Fabris, Alessandro, Messina, Stefano, Silvello, Gianmaria, e Susto, Gian Antonio. 2022. "Algo-set di dati sull'equità ritmica: la storia finora". arXiv prestampa *arXiv:2202.01711*.
- 565 Denton, Emily, Hanna, Alex, Amironesei, Razvan, Smart, Andrew e Nicole, Hilary. 2021. "Sulla genealogia dei set di dati di machine learning: una storia critica di imangenet". *Big Data e Società*, 8(2):20539517211035955.
- 566 Raji, Inioluwa Deborah, Bender, Emily M, Paullada, Amandalynne, Denton, Emily e Hanna, Alex. 2021. "Ai e il punto di riferimento di tutto nel mondo intero". arXiv prestampa *arXiv:2111.15366*.
- 567 Hand, David J.. 2010. Teoria e pratica della misurazione: il mondo attraverso la quantificazione. Wiley.
- 568 Hand, David J. 2016. Misurazione: una breve introduzione. Stampa dell'Università di Oxford.
- 569 Bandalos, Deborah L. 2018. Teoria della misurazione e applicazioni per le scienze sociali. Guilford Pubblicazioni.
- 570 Liao, Thomas, Taori, Rohan, Raji, Inioluwa Deborah e Schmidt, Ludwig. 2021. "Stiamo ancora imparando? una metarevisione degli errori di valutazione nell'ambito del machine learning". Nella trentacinquesima conferenza sui set di dati e sui benchmark dei sistemi di elaborazione delle informazioni neurali (round 2).