Paolo Benanti: **AI RESPONSABILI: IMPARIAMO DAL PASSATO** (blog, 13 febbraio 2020)

Imparare dal passato per creare un'AI responsabile significa confrontarsi con una raccolta di casi d'uso almeno controversi e spesso non etici che hanno caratterizzato lo scenario del mondo delle AI. A seguire una raccolta per temi.

La seguente raccolta di link (tutto materiale originale in inglese) serve come riferimento per ingegneri, data scientist, eticisti, policy makers e altri che prendono decisioni sulla costruzione di soluzioni tecnologiche per problemi del mondo reale. Speriamo che ciò ci aiuti a evitare di ripetere gli errori del passato informando la progettazione di nuovi sistemi o la decisione di non creare affatto una soluzione tecnologica.

1. PROBLEMI DI EQUITÀ E GIUSTIZIA

Prestito e approvazione del credito

- Reclami di pregiudizio di genere contro la Apple Card segnalano un lato oscuro a Fintech
- Esplorazione della discriminazione razziale nei prestiti ipotecari: un appello per una maggiore trasparenza
- DFS pubblica una guida agli assicuratori vita sull'uso di "dati esterni" nelle decisioni sulla sottoscrizione

Assunzioni

- Amazon elimina lo strumento di reclutamento segreto dell'IA che ha mostrato pregiudizi contro le donne
- Discriminazione dell'occupazione automatizzata
- L'aiuto desiderato: un esame degli algoritmi di assunzione, equità e distorsione
- Tutti i modi in cui gli algoritmi di assunzione possono introdurre distorsioni
- Distorsione da mitigazione nelle assunzioni algoritmiche: valutazione di reclami e pratiche
- Help Wanted - A Examination of Hiring Algorithms, Equity and Bias
- Ricercato: la "babysitter perfetta". Deve passare la scansione AI per rispetto e attitudine.

Valutazione dei dipendenti

- La Houston Houston deve affrontare una causa per la valutazione degli insegnanti
- In che modo Amazon rintraccia e licenzia automaticamente i magazzinieri per la "produttività"

Valutazione del rischio pre-processuale e condanna penale

- Distorsione da macchina
- Come abbiamo analizzato l'algoritmo di recidiva COMPAS
- Archivio GitHub per analisi COMPAS
- Riesci a rendere l'IA più giusta di un giudice? Gioca al nostro algoritmo di gioco in tribunale

Casi d'uso di polizia preventiva e altre forze dell'ordine

- Dati sporchi, previsioni errate: in che modo le violazioni dei diritti civili incidono sui dati della polizia, sui sistemi di polizia predittiva e sulla giustizia
- Il riconoscimento facciale di Amazon ha eguagliato 28 membri del Congresso con foto segnaletiche
- The Perpetual Line-Up - La polizia non regolamentata deve essere riconosciuta in America
- Stuck in a Pattern: prime prove di "polizia preventiva" e diritti civili
- Lo studio mostra PredPol, strumento di previsione del crimine che amplifica le attività di polizia razzialmente distorte
- Apprendimento automatico criminale
- The Liar's Walk - Rilevare l'inganno con andatura e gesti

- Lo studio federale conferma il pregiudizio razziale di molti sistemi di riconoscimento facciale, mette in dubbio il loro uso in espansione

Ammissioni

- British Medical Journal: una macchia sulla professione

Scelta della scuola

- Il software personalizzato aiuta le città a gestire la scelta della scuola

Rilevazione vocale

- Oh caro ... I modelli di intelligenza artificiale usati per segnalare i discorsi di odio online sono, ehm, razzisti contro i neri
- Il rischio di parzialità razziale nel rilevamento del discorso dell'odio
- Tossicità e tono non sono la stessa cosa: analizzare la nuova API di Google sulla tossicità, PerspectiveAPI.
- La voce è la prossima grande piattaforma, a meno che tu non abbia un accento
- Il riconoscimento vocale di Google ha un orientamento al genere

Etichettatura delle immagini e riconoscimento facciale

- Google Foto ha identificato due neri come "gorilla"
- Quando si tratta di gorilla, Google Foto rimane cieco
- L'app virale selfie ImageNet Roulette sembrava divertente - fino a quando non mi ha definito un insulto razzista
- Google sta studiando il motivo per cui ha studiato il riconoscimento facciale dei senzatetto "dalla pelle scura"
- Sfumature di genere: disparità di precisione intersezionale nella classificazione commerciale di genere
- Le macchine insegnate dalle foto imparano una visione sessista delle donne
- Gli inquilini hanno lanciato l'allarme per il riconoscimento facciale nei loro edifici. I legislatori stanno ascoltando.

Benefici pubblici e salute

- Un algoritmo di assistenza sanitaria che colpisce milioni è distorto contro i pazienti di colore
- Cosa succede quando un algoritmo taglia la tua assistenza sanitaria
- La Cina sa come portare via la tua assicurazione sanitaria
- Predire il futuro: una prospettiva critica sull'uso dell'analisi predittiva nel benessere dei minori
- Non esiste una soluzione rapida per trovare la distorsione razziale negli algoritmi sanitari

Annunci

- Discriminazione nella pubblicazione di annunci online
- Sondare il lato oscuro del sistema di targeting degli annunci di Google
- Facebook si impegna nella discriminazione degli alloggi con le sue pratiche pubblicitarie, secondo gli Stati Uniti
- Le offerte di lavoro di Facebook sollevano preoccupazioni circa la discriminazione in base all'età
- Gli annunci di Facebook possono ancora discriminare donne e lavoratori anziani, nonostante una soluzione per i diritti civili
- Lo studio mostra che alle donne è meno probabile che vengano mostrati annunci per lavori ben pagati su Google
- Algoritmi che "non vedono il colore": confrontare i pregiudizi nei sosia e nei segmenti di pubblico speciali

Ricerca

- Algoritmi di oppressione: come i motori di ricerca rafforzano il razzismo

- Il pregiudizio esiste già nei risultati dei motori di ricerca e peggiorerà solo
- La verità nelle immagini: ciò che le ricerche di immagini di Google ci dicono sulla disuguaglianza sul lavoro

Traduzioni

- Google Translate potrebbe avere un problema di genere

Selezione della giuria

- La giuria dei big data

Incontri

- Il caffè incontra il bagel: il sito di incontri online che ti aiuta a eliminare i brividi
- I pregiudizi che forniamo agli algoritmi di Tinder
- Riprogetta le app di appuntamenti per ridurre il pregiudizio razziale, raccomanda lo studio

Incorporamenti di parole

Gli incorporamenti di parole possono influire su molte delle categorie precedenti tramite le applicazioni che le utilizzano.

- L'uomo è programmatore di computer come la donna è casalinga? Debiasing Word Embeddings

Manipolazione

- Quando si traccia una linea è difficile

2. PROBLEMI DI SICUREZZA

Auto a guida autonoma

- Ricordi l'auto a guida autonoma Uber che ha ucciso una donna che attraversava la strada? L'intelligenza artificiale non aveva idea di jaywalker

AI armate

- Protesta dei dipendenti di Google: ora Google interrompe il progetto AI del drone del Pentagono
- Google vuole fare affari con i militari, molti dei suoi dipendenti non lo fanno

3. PROBLEMI DELLA SANITÀ E DELLA MEDICINA

Interpretazione del modello in medicina

- Modelli intelligenti per HealthCare: previsione del rischio di polmonite e riammissione di 30 giorni in ospedale mostrano l'importanza dell'interpretazione del modello per tali decisioni critiche.
- Rich Caruana: gli amici non lasciano che gli amici rilascino modelli di scatole nere in medicina
- IBM ha lanciato il suo supercomputer Watson come una rivoluzione nella cura del cancro. Non è affatto vicino
- Valutazione internazionale di un sistema di IA per lo screening del cancro al seno - Questo thread esamina i problemi con l'impostazione del problema.

4. PROBLEMI DI PRIVACY

Attacchi alla privacy basati su Machine Learning

- Attacchi alla privacy sui modelli di apprendimento automatico

Prestiti bancari

- Il nuovo mondo del prestito, post-demonetizzazione
- Debito perpetuo nella savana di silicio

Posti di lavoro

- Donna licenziata dopo aver disabilitato l'app di lavoro che monitorava i suoi movimenti 24/7
- Sorveglianza illimitata dei lavoratori

Tecnologia carceraria

- La società di tecnologia carceraria viene interrogata per il mantenimento di "impronte vocali" di persone ritenute innocenti

Dati sulla posizione

- Dodici milioni di telefoni, un set di dati, Zero Privacy fa luce sulla privacy dei dati (o sulla loro mancanza). Gli stessi dati possono essere utilizzati anche per ML.
- Gli inquilini hanno lanciato l'allarme per il riconoscimento facciale nei loro edifici. I legislatori stanno ascoltando.

Social media e incontri

- Lo studio di OkCupid rivela i pericoli della scienza dei Big Data
- "We Are the Product": reazioni pubbliche alla condivisione di dati online e controversie sulla privacy nei media

Anonimizzazione di base come misura insufficiente

- I dati "anonimi" in realtà non lo sono — ed ecco perché no

Salute

- Gli assicuratori sanitari stanno aspirando i dettagli su di te e potrebbero aumentare le tue tariffe
- Come i tuoi dati medici alimentano un'industria nascosta di svariati miliardi di dollari
- Le offerte farmaceutiche di 23andMe sono sempre state il piano
- Se vuoi un'assicurazione sulla vita, pensa due volte prima di fare un test genetico
- La start-up medica ha invitato milioni di pazienti a scrivere recensioni che potrebbero non realizzare sono pubbliche. Alcuni sono espliciti.
- Help desk: le tue cartelle cliniche possono diventare marketing? Indaghiamo il "portale paziente" sospetto di un lettore.
- L'app per la gravidanza sta condividendo i tuoi dati intimi con il tuo capo?
- Crisi dei dati: chi detiene la cartella clinica?
- Questo tampone Bluetooth è la cosa più intelligente che puoi mettere nella tua vagina non menzionando le preoccupazioni sulla privacy di un tale dispositivo. Questo commento Twitter aggiunge il commento necessario.

Antiriciclaggio

- Fidarsi dell'apprendimento automatico nel settore del riciclaggio di denaro: un approccio basato sul rischio

5. RISORSE GENERALI SULL'AI RESPONSABILE

Molti dei libri e degli articoli in quest'area coprono una vasta gamma di argomenti. Di seguito è riportato un elenco di alcuni di essi, in ordine alfabetico per titolo:

- Un giuramento di Ippocrate per i professionisti dell'intelligenza artificiale di Oren Etzioni
- Algorithms of Oppression - In che modo i motori di ricerca rafforzano il razzismo di Safiya Umoja Noble
- Unintelligence artificiale: come i computer fraintendono il mondo di Meredith Broussard
- Automatizzare la disuguaglianza: come gli strumenti ad alta tecnologia definiscono, sorvegliano e puniscono i poveri di Virginia Eubanks
- Disparate Impact di Big Data di Solon Barocas e Andrew D. Selbst
- Equità e astrazione nei sistemi sociotecnici di Andrew D. Selbst , danah boyd , Sorelle Friedler , Suresh Venkatasubramanian , Janet Vertesi

- Equità e apprendimento automatico - Limitazioni e opportunità di Solon Barocas , Moritz Hardt , Arvind Narayanan
- Come sto combattendo il pregiudizio negli algoritmi di Joy Buolamwini
- Apprendimento automatico interpretabile di Christoph Molnar
- Curriculum etico tecnico di Casey Fiesler
- Weapons of Math Destruction di Cathy O'Neil