Pristranskost v umetni inteligenci

...in njen vpliv na človekove pravice

Matej Kovačič, Alenka Guček, Tanja Zdolšek Draksler

https://ircai.org

Institut "Jožef Stefan" Ljubljana, Slovenija

IRC AI
International Research Centre of Artificial Inteligence under the auspices of UNESCO

Odsek za Umetno Inteligenco

10. dnevi prava zasebnosti in varovanja informacij

Ptuj, 18. in 19. april 2024

Pristranskost

Pristranskost (angl. bias): vnaprejšnje predstave, predsodki ali nepoštene preference, ki izvirajo iz osebnih izkušenj, kulturnih vplivov ali družbenih pogojev.

Pristranskost lahko pomembno vpliva na to, kako posamezniki dojemajo, presojajo in se odločajo v različnih situacijah.

Pristranskost v UI

Do pristranskosti v UI pride, ko algoritmi ustvarijo sistematično napačne rezultate zaradi pristranskih predpostavk med razvojem ali v podatkih za učenje modelov.

Pristranskost v umetni inteligenci (UI) se nanaša na situacije, ko UI sistemi določene skupine posameznikov obravnavajo nepravično ali diskriminatorno.

UI rešitev je pristranska, če vrača izkrivljene rezultate, predvsem v primerih, ki krepijo škodljiva prepričanja ali podpirajo predsodke do posameznikov ali skupin. Taka UI rešitev tako ni v skladu s pozitivnimi človeškimi vrednotami.

Pristranskost v UI

Pristranskost se lahko kaže na različne načine, do nje pa lahko pride na različnih stopnjah razvoja UI rešitve:

- med zbiranjem in predobdelavo podatkov,
- pri izbiri vhodnih parametrov za model,
- med gradnjo modela,
- med vrednotenjem modela in končnim uvajanjem UI rešitve v prakso.

Težava pa je tudi v tem, da se lahko celo manjša začetna pristranskost med procesom strojnega učenja poveča, kar na koncu zastrupi celoten sistem in privede do resne pristranskosti celotne UI rešitve.

Primer #1: medicina

Skupina raziskovalcev na urgentnem oddelku bolnišnice Duke University je začela razvijati algoritem za pomoč pri napovedovanju otroške sepse. Ker so se zavedali problema pristranskosti, so v odpravo le-te vložili veliko truda.

A po skoraj treh letih so ugotovili, da so zdravniki na podlagi ocene UI rešitve, potrebovali dlje, da so naročali krvne preiskave za latinoameriške otroke, ki so jim na koncu diagnosticirali sepso, kot pa za belopolte otroke.

Eden izmed identificiranih možnih vzrokov za to je bil, da je potreba po tolmačih upočasnila proces zbiranja zdravstvenih podatkov ter naročanje krvnih preiskav.

Ta zamuda pa je umetno inteligenco napačno naučila, da otroci latinsko ameriškega porekla razvijejo sepso počasneje kot drugi otroci.

Primer #2: odkrivanje goljufij

Leta 2023 so novinarji časnika Guardian razkrili, da britansko notranje ministrstvo za označevanje navideznih porok uporablja UI rešitev, ki pa je bila pristranska do ljudi določenih narodnosti. Notranja ocena notranjega ministrstva je pokazala, da orodje nesorazmerno bolj kot goljufe označuje ljudi iz Albanije, Grčije, Romunije in Bolgarije.

Pristransko UI rešitev za odkrivanje goljufij pri dodeljevanju denarnih nadomestil pa je uporabljalo tudi ministrstvo za delo. Algoritem je kot potencialne goljufe napačno označil nesorazmerno veliko Bolgarov, ki so posledično nato izgubili denarna nadomestila.

Odgovor britanskih ministrstev je bil, da je celoten sistem vseeno pošten, saj UI daje samo *priporočila*, končne odločitve pa sprejemajo ljudje. V resnici pa se praksi uradniki zelo *zanašajo* na odločitve algoritmov.

Pristranski algoritmi tako vodijo do pristranskih končnih odločitev. Ljudje, na katere te odločitve vplivajo pa sploh ne izvedo, da je odločitev temeljila na pristranski umetni inteligenci.

Primer #3: otroški dodatki

Pristransko UI rešitev za ocenjevanje tveganja za goljufije z otroškimi dodatki, so uporabljali tudi na Nizozemskem, kjer so se davčni organi osredotočali na ljudi s turško ali maroško narodnostjo, med dejavniki tveganja pa so bili tudi dvojno državljanstvo in nizki dohodki.

Oblasti so družine kaznovale zgolj zaradi suma goljufije na podlagi indikatorjev tveganja sistema.

Rezultati so bili katastrofalni. Nizozemska davčna uprava je od ljudi, ki so bili označeni kot goljufi, zahtevala vračilo dodatkov za varstvo otrok, več deset tisoč družin pa je bilo zaradi previsokih dolgov do davčne agencije pahnjenih v revščino.

Nekatere žrtve so naredile samomor in več kot tisoč otrok je bilo poslanih v rejništvo.

Primer #4: prepoznava obrazov

Več raziskav je pokazalo, da se številni algoritmi za prepoznavanje obraza slabo obnesejo pri prepoznavanju ljudi, ki niso belci. Raziskovalci so ugotovili, da je to posledica dveh dejavnikov: v naborih podatkov za usposabljanje algoritmov je premalo obrazov temnopoltih, in drugič - ti sistemi pogosto povečajo lastne pristranskosti policistov.

To pa lahko vodi do večjega rasnega profiliranja in več lažnih aretacij. Netočna identifikacija namreč poveča verjetnost zgrešenih aretacij.

Programska oprema seveda ne aretira ljudi. Vendar pa nakaže, kdo so potencialni osumljenci, policisti pa se nato odločijo, koga bodo aretirali.

Toda ljudje pogosto verjamejo, da je umetna inteligenca nezmotljiva, in ne dvomijo v rezultate.

Primer #4: prepoznava obrazov

Leta 2020 je bil Robert Williams aretiran zaradi domnevne kraje več tisoč dolarjev vrednih ur. Detroitska policija je uporabila algoritem umetne inteligence, ki je povezal video iz nadzorne kamere s fotografijo iz Williamsovega vozniškega dovoljenja.

Policija ni zbrala nobenih drugih dokazov, kot so identifikacija očividca, podatki o lokaciji mobilnega telefona ali prstni odtis.

Edini dokaz je bila slika iz videonadzorne kamere in identifikacija z algoritmom AI.

Williamsa so nato aretirali ter ga pridržali 18 ur. Njegova aretacija je prvi dokumentiran primer, ko je bil nekdo neupravičeno pridržan na podlagi tehnologije za prepoznavanje obraza.

Spopadanje s pristranskostjo

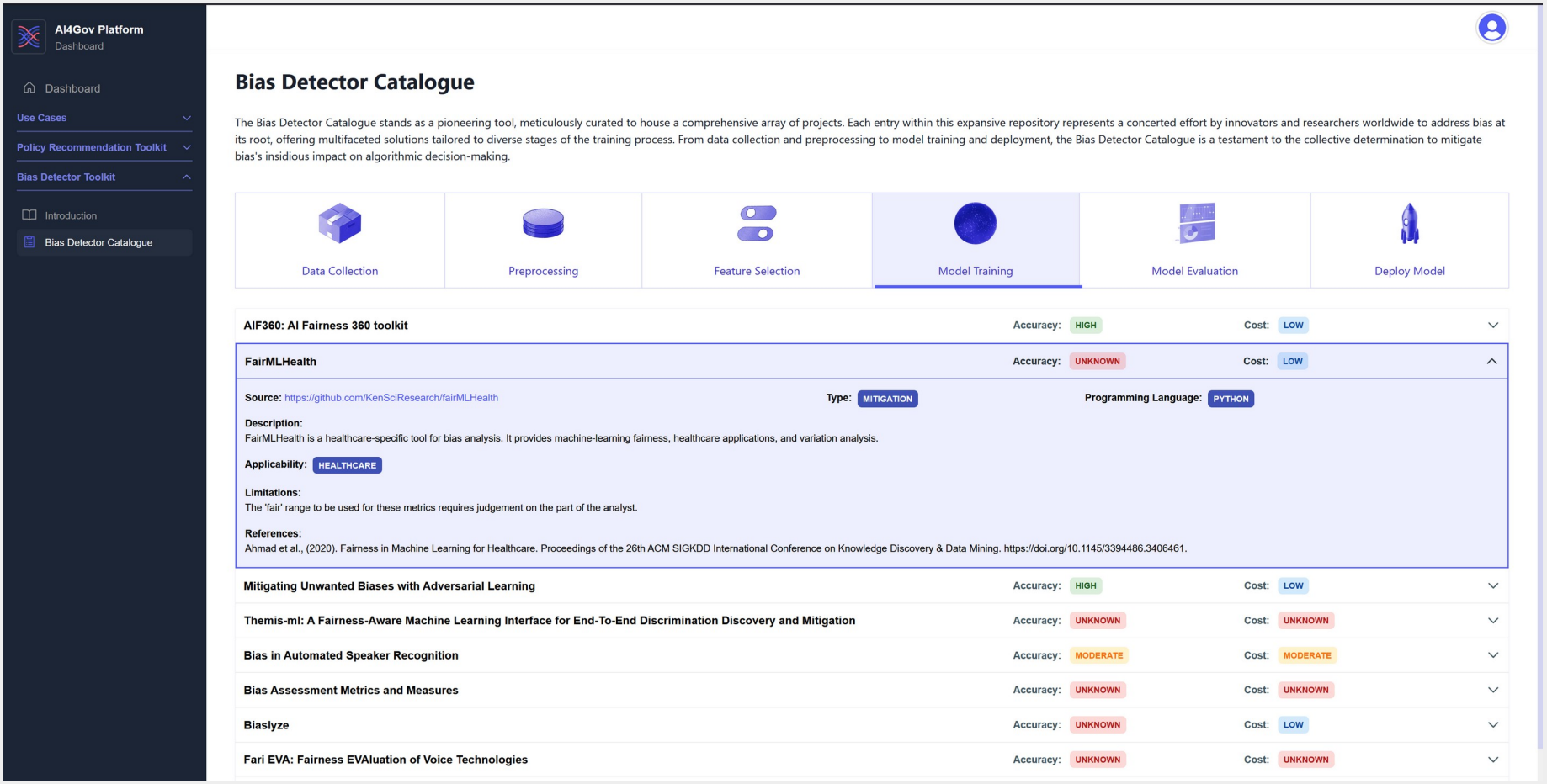
Obstajajo različne **strategije** (adversarial debiasing, reweighing ter ensemble adversarial training,...) in **orodja** (IBM-ov AI Fairness 360, Googlovo What-If orodje, LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations) za blaženje pristranskosti, ki so usmerjene na različne stopnje razvoja sistemov UI.

AI4Gov
Trusted AI for Transparent Public Governance
fostering Democratic Values

V okviru projekta AI4Gov razvijamo **integracijo** teh **tehnik** za blaženje pristranskosti, **XAI** (explainable artificial intelligence) **orodij** ter **izobraževalnih gradiv** iz področja pristranskosti v AI.

Spopadanje s pristranskostjo

Pripravljamo tim. **Bias Detector katalog**, ki bo na pregleden način predstavil različna orodja za analizo in blaženje pristranskosti v UI. Orodja so kategorizirana tudi glede na funkcionalnosti, področje uporabe in performančnost.

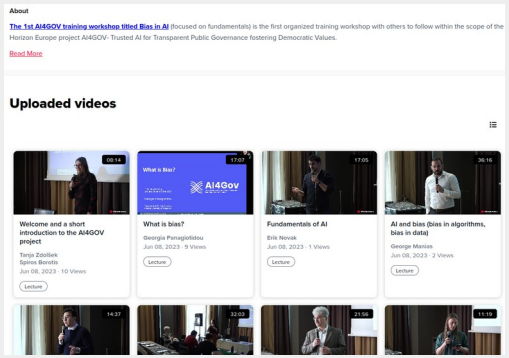


Spopadanje s pristranskostjo

Eden izmed ciljev AI4Gov projekta je tudi priprava različnih izobraževanj na to tematiko.

MOOC: Trustworthy and democratic AI – Fundamentals (Zaupanja vredna umetna inteligenca, ki spodbuja demokratične vrednote)

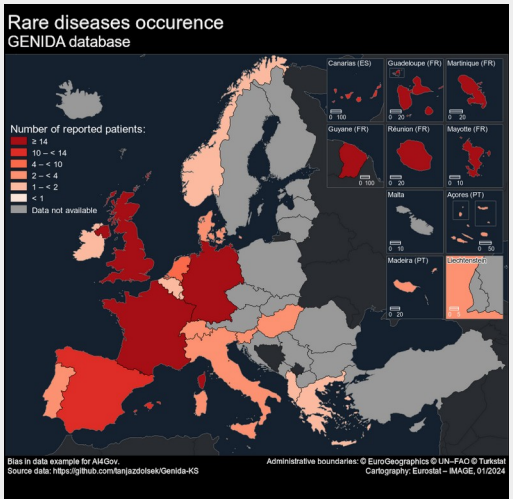
- Stopnja: začetnik
- Trajanje: 10 ur
- Prosto dostopno na portalu OpenLearnCreate (https://www.open.edu/openlearncreate/enrol/index.php?id=11669)
- Pripravljata se tudi slovenska in španska verzija.

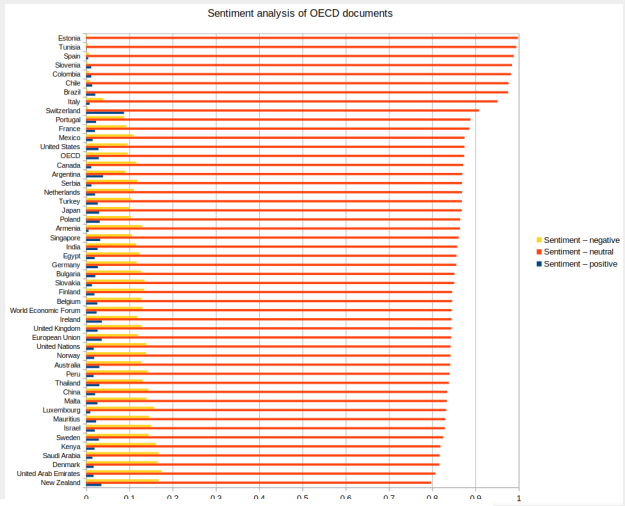


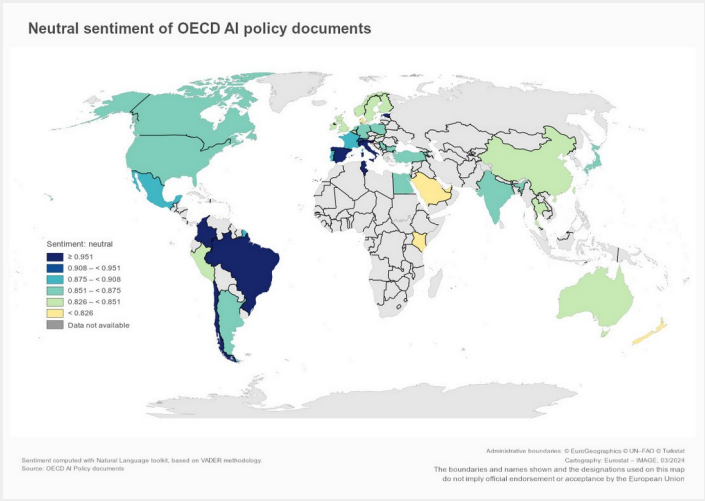
https://videolectures.net/AI4GOVtraining2023_ljubljana/

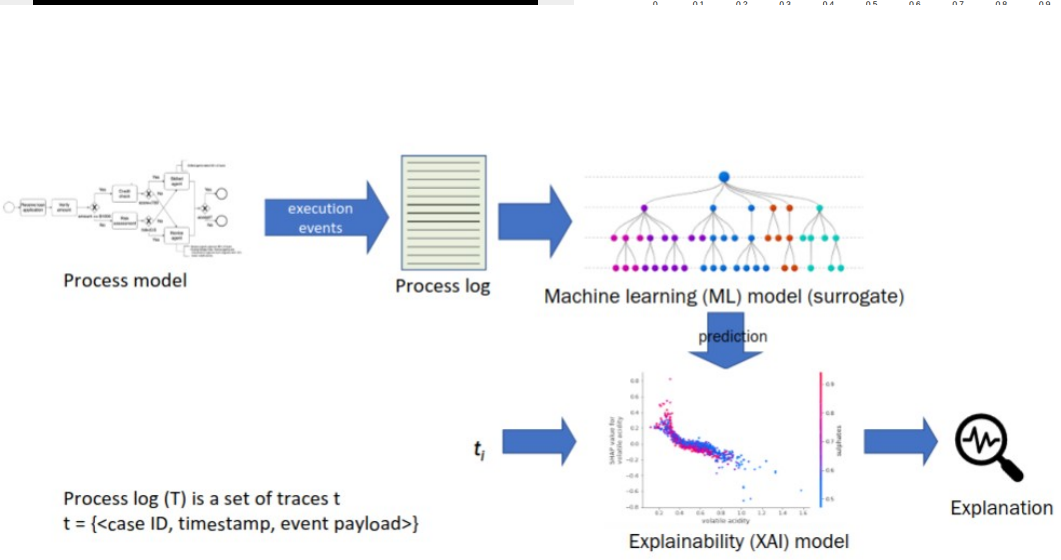
Spopadanje s pristranskostjo

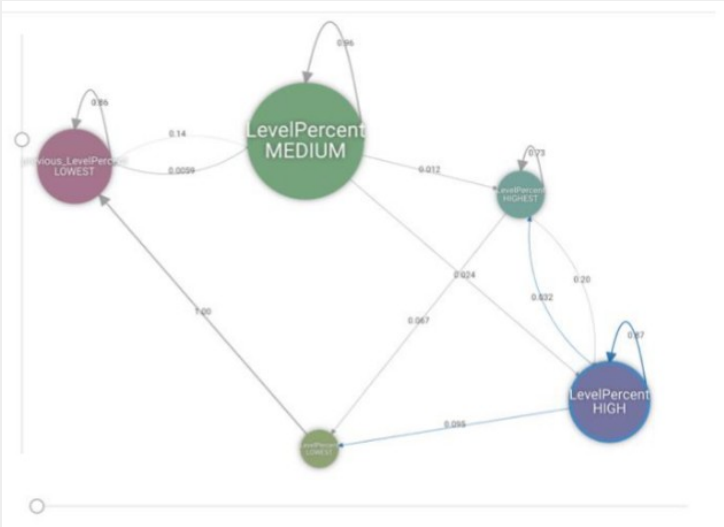
Pripravljamo pa tudi tri pilotne projekte in testiramo različne metode in pristope za analize pristranskosti.











Vprašanja?

AI4Gov

Trusted AI for Transparent Public Governance
fostering Democratic Values

https://ai4gov-project.eu/