

Datum: 31. 3. 2025

Pristranskost algoritmov

Umetna inteligenca in algoritmi vse pogostejše pomagajo pri odločanju v šolah, podjetjih, zdravstvu, javni upravi, spletnih storitvah in varnosti. Če podatki niso kakovostni, uravnoteženi ali pravično uporabljeni, lahko sistem sprejme nepravične ali škodljive odločitve.





Kaj je algoritem?

- Zaporedje pravil ali navodil, po katerih računalnik reši določeno nalogo.
- Lahko razvršča podatke in išče vzorce.
- Priporoča vsebine in ocenjuje tveganja.
- Prepoznavna slike in obdeluje besedilo.
- Pomaga pri odločanju.
- Algoritem ni nujno nevtralen, saj je odvisen od podatkov in ciljev, ki jih določijo ljudje.



Kaj pomeni pristranskost?

- ✦ Pristranskost pomeni, da sistem neko osebo, skupino ali možnost obravnava nepošteno.
- ✦ Kaže se lahko kot neupravičena prednost ali slabša obravnava.
- ✦ Vključuje napačno razvrščanje ali izključevanje.
- ✦ Lahko vodi v stereotipno sklepanje.
- ✦ Rezultat je nepravilna odločitev.
- ✦ Pri umetni inteligenci pristranskost pogosto ni namerna, vendar ima lahko zelo resnične posledice.

Zakaj je pristranskost algoritmov pomembna?

- ✦ Vpliva na zaposlitev in dostop do izobraževanja.
- ✦ Vpliva na zdravstveno obravnavo in kreditno sposobnost.
- ✦ Dotika se zavarovanj in varnostnih ocen.
- ✦ Upliva na priporočila na spletu in dostop do storitev.
- ✦ Nepravična odločitev algoritma lahko povzroči škodo posamezniku.
- ✦ Pomembne življenjske odločitve so pod vplivom algoritmov.

Umetna inteligenca se uči iz podatkov

- ➔ Sistem prejme veliko primerov.
- ➔ V podatkih išče vzorce.
- ➔ Iz vzorcev oblikuje model.
- ➔ Model nato uporablja pri novih primerih.
- ➔ Če so podatki slabi ali neuravnoteženi, se lahko sistem nauči napačnih vzorcev.



Slabi podatki pomenijo slabe odločitve

- Slabi podatki so lahko netočni ali zastareli.
- Pogosto so nepopolni ali napačno označeni.
- Lahko so zbrani iz omejenega ali nereprezentativnega okolja.
- Vsebujejo veliko napak.
- Kakovost podatkov močno vpliva na kakovost odločitev umetne inteligence.
- Umetna inteligenca ni boljša od podatkov, na katerih temelji.



Neuravnoteženi podatki

- ➔ Nekatere skupine so v podatkih dobro zastopane, druge pa premalo.
- ➔ Primeri vključujejo več podatkov o eni starostni skupini ali jeziku.
- ➔ Premalo primerov iz določenega okolja ali države.
- ➔ Premalo različnih obrazov pri učenju prepoznavanja.
- ➔ Če sistem neke skupine ne vidi dovolj pogosto, jo lahko slabše prepozna ali napačno obravnava.

Zgodovinska pristranskost

- ➔ Podatki odražajo pretekle nepravilnosti ali neenakosti, kot je neenako zaposlovanje.
- ➔ Določene skupine so imele manj možnosti.
- ➔ Podatki odražajo stare predsodke.
- ➔ Če se umetna inteligenca uči iz nepravilne preteklosti, lahko nepravilnost prenese v prihodnost.

Pristranskost pri zbiranju podatkov

- ➔ Vprašanja glede tega, kdo je bil vključen in kdo izključen.
- ➔ Kako so bili podatki zbrani in ali so dovolj raznoliki.
- ➔ Ali so podatki zbrani zakonito in etično.
- ➔ Ali podatki prikazujejo resnično stanje.
- ➔ Način zbiranja podatkov lahko določi, komu bo sistem kasneje bolj ali manj ustrezal.

Pristranskost pri označevanju podatkov

- ➔ Umetna inteligenca pogosto potrebuje označene podatke, npr. slika mačke ali psa.
- ➔ Sporočilo je označeno kot nevarno ali varno.
- ➔ Kandidat je označen kot primeren ali neprimeren.
- ➔ Če so oznake napačne ali pristranske, se sistem nauči napačnih povezav.

Pristranskost pri oblikovanju sistema

- ➔ Razvijalci lahko izberejo napačen cilj ali sistem optimizira samo en rezultat.
- ➔ Ne preverja se vpliv na različne skupine.
- ➔ Ni dovolj testiranja.
- ➔ Ni vključeno dovolj različnih pogledov.
- ➔ Tehnična rešitev vedno odraža tudi odločitve ljudi, ki so jo zasnovali.

Primer pri zaposlovanju

- ➔ Sistem se uči iz preteklih zaposlitev in daje prednost pogostim profilom.
- ➔ Lahko spregleda dobre kandidate ali kaznuje netipične življenjepise.
- ➔ Napačno razume prekinitve v karieri.
- ➔ Pri zaposlovanju mora biti AI pomoč, ne pa samostojni končni odločevalec.

Primer pri obraznem prepoznavanju

➔ Sistemi so lahko manj natančni za nekatere skupine ljudi.

➔ Razlogi vključujejo premalo raznolikih fotografij, različno osvetlitev ali starosti.

➔ Neuravnoteženi učni podatki prispevajo k težavam.

➔ Napačna identifikacija lahko povzroči resne posledice, zlasti pri varnosti.

Primer pri kreditih in zavarovanjih

- ➔ Algoritmi lahko napačno ocenijo posameznika ali tveganje.
- ➔ Uporabijo posredne podatke in kaznujejo ljudi iz določenega okolja.
- ➔ Ne upoštevajo posebnih okoliščin in utrjujejo socialne razlike.
- ➔ Odločitev o kreditu ali zavarovanju bistveno vpliva na življenje posameznika.
- ➔ Mora biti poštena in razložljiva.

Primer pri zdravstvu

- ➔ Sistem je naučen na podatkih ene skupine pacientov in slabše deluje pri drugih.
- ➔ Ne upošteva posebnih okoliščin ali napačno razvrsti tveganje.
- ➔ Zdravnik preveč zaupa rezultatu.
- ➔ V zdravstvu mora biti uporaba umetne inteligence posebej previdna, ker lahko vpliva na zdravje in varnost.

Primer pri izobraževanju

- ➔ Algoritmi se uporabljajo za priporočanje učnih vsebin in ocenjevanje napredka.
- ➔ Odkrivajo težave pri učenju in razvrščajo učence.
- ➔ Omogočajo samodejno ocenjevanje nalog.
- ➔ Če sistem napačno oceni učenca, lahko vpliva na njegovo samozavest, podporo in izobraževalne možnosti.

Posredna diskriminacija

- ➔ Sistem ne uporablja občutljivega podatka neposredno, ampak drug podatek s podobnim učinkom.
- ➔ Primeri vključujejo poštno številko, šolo ali delovno zgodovino.
- ➔ Tudi način pisanja, lokacija ali spletno vedenje so lahko posredni podatki.
- ➔ Tudi če sistem ne uporablja neposredno osebnih značilnosti, lahko prek drugih podatkov pride do nepravičnih rezultatov.



Črna skrinjica



Nekateri sistemi umetne
intelligence so težko razumljivi.



Vemo, kaj vnesemo, in vidimo rezultat, a ne
razumemo, kako je sistem prišel do odločitve.



Pri pomembnih odločitvah je nevarno, če
nihče ne zna pojasniti, zakaj je sistem odločil
tako, kot je.

Preglednost odločanja

- ➔ Uporabniki morajo razumeti, kdaj in kako se uporablja algoritem.
- ➔ Pomembno je vedeti, kateri podatki so uporabljeni in kakšen je namen sistema.
- ➔ Kdo je odgovoren in kako je mogoče odločitev preveriti.
- ➔ Kako se oseba lahko pritoži.
- ➔ Ljudje morajo imeti možnost razumeti in izpodbijati odločitve, ki vplivajo na njihovo življenje.

Razložljivost

- ➔ Sistem ali organizacija mora pojasniti rezultat na razumljiv način.
- ➔ Pojasniti mora, kateri dejavniki so vplivali na odločitev in zakaj je bil rezultat takšen.
- ➔ Ali je bil uporabljen pravilen podatek in ali je možen popravek.
- ➔ Kdo lahko odločitev pregleda.
- ➔ Razložljivost krepi zaupanje in omogoča popraviljanje napak.

Človeški nadzor

- ➔ Pomembnih odločitev ne prepustimo popolnoma algoritmu.
- ➔ Potreben je pri zaposlovanju in zdravstvu.
- ➔ Tudi v javni upravi, izobraževanju in varnosti.
- ➔ Ključen pri kreditih in zavarovanjih.
- ➔ Človek mora imeti možnost preveriti rezultat in prevzeti odgovornost za končno odločitev.

Testiranje algoritmov

- Pred uporabo je treba algoritme testirati.
- Testiranje preverja natančnost in pravičnost.
- Preverja delovanje pri različnih skupinah in odkriva napake.
- Zaznava neželene učinke in zagotavlja varnost.
- Preverja razložljivost sistema.
- Algoritem, ki dobro deluje v povprečju, lahko še vedno slabo deluje za določeno skupino ljudi.

Redno spremljanje po uvedbi

- ➔ Tudi po uvedbi sistema je treba spremljati njegovo delovanje.
- ➔ Podatki in družba se spreminjajo.
- ➔ Uporabniki sistem uporabljajo drugače in pojavijo se lahko nove napake.
- ➔ Sistem lahko začne delovati slabše.
- ➔ Pravičnost algoritma ni enkraten pregled, ampak stalni proces.

Raznolike razvojne ekipe

- ➔ Raznolike ekipe lahko bolje prepoznajo tveganja pristranskosti.
- ➔ Pomagajo pri širšem razumevanju uporabnikov in prepoznavanju slepih peg.
- ➔ Pripomorejo k boljšemu oblikovanju vprašanj in pravičnejšemu testiranju.
- ➔ Zagotavljajo odgovorno uporabo sistema.
- ➔ Če sistem razvija ozka skupina ljudi, lahko spregleda izkušnje drugih skupin.

Varstvo osebnih podatkov

- ➔ Pomembno je zakonito zbiranje in minimalna uporaba podatkov.
- ➔ Jasni namen in varnost podatkov sta ključna.
- ➔ Omejen dostop in spoštovanje pravic posameznikov.
- ➔ Preprečevanje zlorab.
- ➔ Pravična umetna inteligenca mora biti tudi zakonita, varna in spoštljiva do zasebnosti.



Odgovornost organizacij

- ➔ Organizacije so odgovorne za izbiro sistema in preverjanje podatkov.
- ➔ Odgovorne so za testiranje in nadzor.
- ➔ Pojasnila uporabnikom in popraviljanje napak.
- ➔ Obravnava pritožb.
- ➔ Organizacija se ne more izgovoriti, da je tako odločil računalnik.

Kako zmanjšamo pristranskost?

- ➔ Uporaba kakovostnih in raznolikih podatkov.
- ➔ Preverjanje neuravnoteženosti in testiranje na različnih skupinah.
- ➔ Zagotavljanje razložljivosti rezultatov in človeškega nadzora.
- ➔ Redno spremljanje in možnost pritožbe.
- ➔ Popolne nepristranskosti je težko doseči, lahko pa zmanjšamo tveganje nepravičnih odločitev.



Vloga uporabnikov

- ✦ Ne zaupamo slepo rezultatom.
- ✦ Preverjamo pomembne odločitve.
- ✦ Vprašamo, kako je rezultat nastal, in opozorimo na napake.
- ✦ Zahtevamo razlago.
- ✦ Uporabljamo lastno presojo.
- ✦ Umetna inteligenca je pomoč, ne nezmotljiv sodnik.

Ključna načela pravične AI

- ➔ **Kakovostni in raznoliki podatki.**
- ➔ **Pravičnost in preglednost.**
- ➔ **Razložljivost in odgovornost.**
- ➔ **Človeški nadzor in varstvo podatkov.**
- ➔ **Ta načela pomagajo zagotoviti, da umetna inteligenca ljudem pomaga, ne pa jih po krivici izključuje.**

Zaključek

- ✦ Pristranskost algoritmov je pomembno vprašanje, ker AI vpliva na ljudi.
- ✦ Pristranskost lahko nastane zaradi slabih ali neuravnoteženih podatkov.
- ✦ Tudi zgodovinske neenakosti in napačno označevanje prispevajo k problemu.
- ✦ Slabo testiranje in pomanjkanje človeškega nadzora so tudi dejavniki.
- ✦ Umetna inteligenca je lahko koristna, vendar mora biti razvita in uporabljena pravično, pregledno ter odgovorno.
- ✦ Le tako lahko zmanjšamo tveganje nepravičnih odločitev.



USPOSABLJAMO
IN PRILAGAJAMO
TRAINING AND ADJUSTMENT



NAČRT ZA
OKREVANJE
IN ODPORNOST



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DELO, DRUŽINO,
SOCIALNE ZADEVE IN ENAKE MOŽNOSTI



Financira
Evropska unija
NextGenerationEU



Umetna inteligenca – napredna Pristranskost algoritmov

Kako lahko umetna inteligenca sprejema nepravične
odločitve zaradi slabih ali neuravnoteženih podatkov

Predavatelj: dr. Miha Dvojmoč

Naročnik: GRIFFIN d.o.o.

Izvajalec: MD svetovanje d.o.o.

Datum: 31. 3. 2025