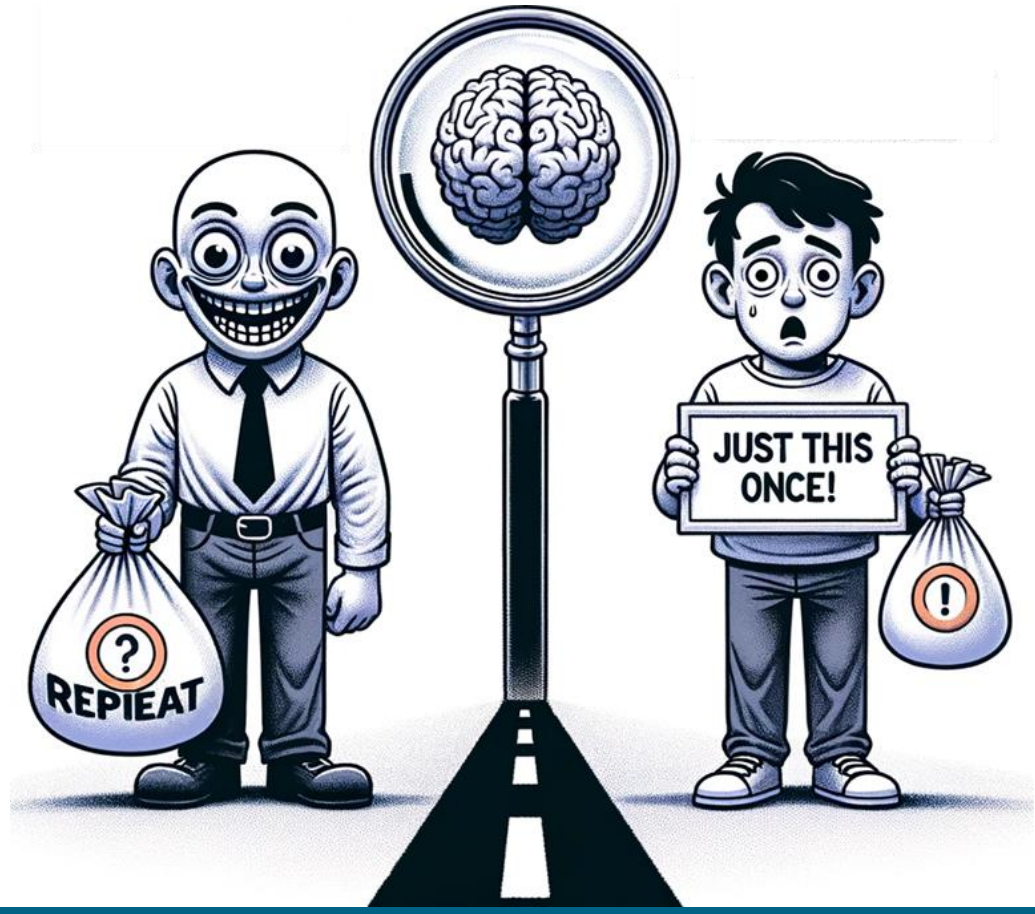
Pravičnost strojnega učenja

Prepoznava problematičnih kršiteljev

- Problem: Vsak, ki stori kaznivo dejanje ni nujno, da je problem za družbo. Lahko, da je kaznivo dejanje bil le enkratni dogodek. Zato želimo redne kršitelje obravnavat drugače kot take, ki so te naredili zgolj izjemoma.

- Kako bi naredili AI model, ki bi ugotavljal, kdo je redni kršitelj in kdo je kršil zgolj izjemoma?
  - Katere podatke bi uporabili?
  - Kaj bi model napovedoval?
  - Kako bi poskrbeli, da bo model deloval enako za vse ljudi?



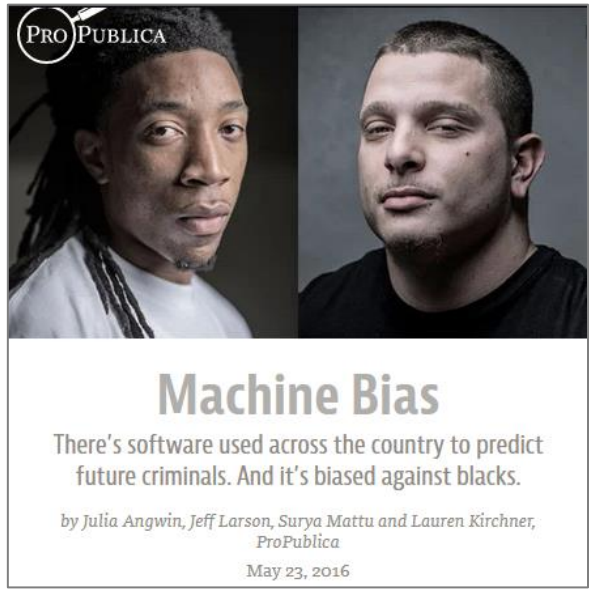
COMPAS

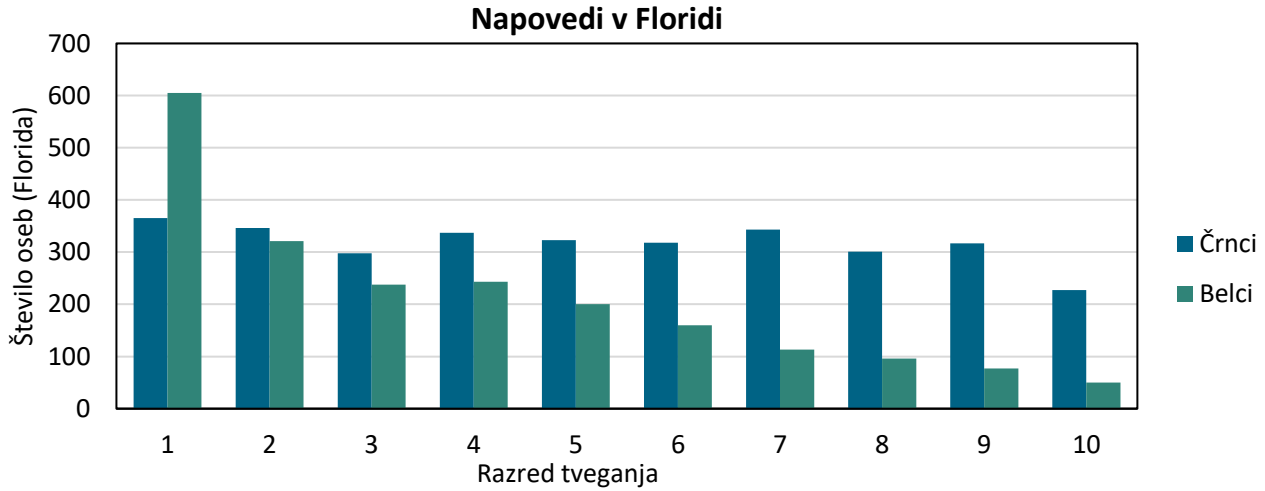
- *Ideja*: Na podlagi zgodovine kršiteljev ugotoviti, če gre za rednega kršitelja.
- *Cilj*: model za izračun tveganja, ki se pa uporablja v dva namena:
  (1) ugotavljanje, če mora biti kršitelj tekom sojenja v **priporu**;
  (2) ugotavljanje, če je obsojenec primeren za **predčasni pogojni izpust**.
- *Podatki*: Zgodovina prejšnjih aretacij, zgodovina zaposlitve, zgodovina naslovov prebivanja, psihiatrične diagnoze, starost, spol, kazenska zgodovina, inteligenca…
  - Klasificirani v enega izmed 10ih razredov glede na tveganje.

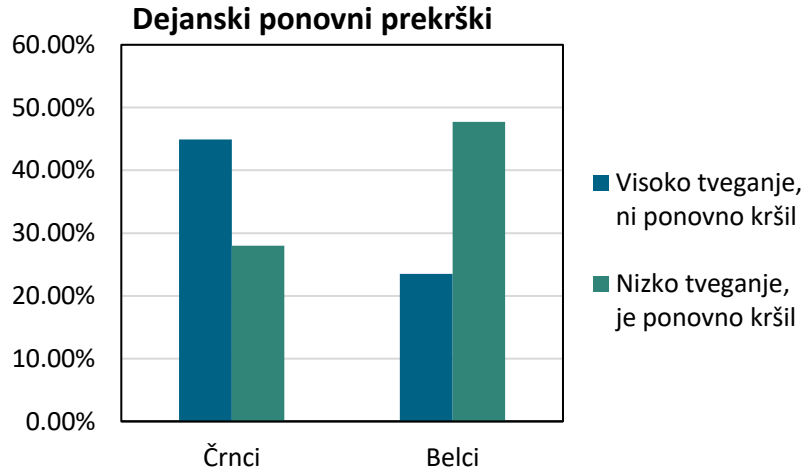
- Poskrbeli so, da je bil COMPAS pravičen na sledeče načine:
  - Ni imel rase kot spremenljivke.
  - Bil je enako točen ne glede na raso (61% točnosti za vse rase).
  - Razredi tveganja so v realnosti pomenili enako ne glede na raso (tveganje 7 je pomenilo enako za vse rase) – kalibriran za raso.

COMPAS

- Leta 2016 *ProPublica* objavi prispevek o analizi sistema.
- Ob pregledu napak (tistih 39%) najdemo sledeče:
  - Za črnce se je večkrat zmotil, da so večje tveganje.
  - Za belce se je večkrat zmotil, da so manjše tveganje.
  - Model ni bil kalibriran za spol (ženske razreda tveganja 5 so enako verjetno ponovile prekršek kot moški tveganja 3).







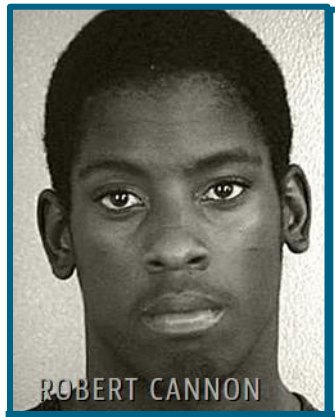
Kriminal: tatvina v trgovini

Predhodni prekrški:

- 1 nasilje v družini s hujšimi posledicami
- 1 velika tatvina
- 1 majhna tatvina
- 1 preprodaja drog

Nizko tveganje **3**

Po izpustu: 1 velika tatvina



Kriminal: tatvina v trgovini

Predhodni prekrški:

- 1 majhna tatvina

Srednje tveganje **6**

Po izpustu: /

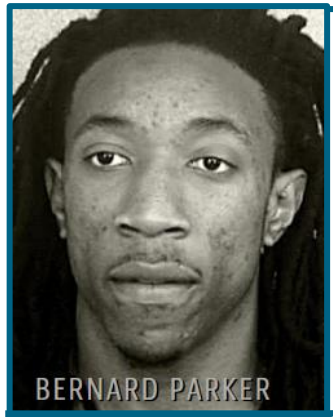
Kriminal: posestvo droge

Predhodni prekrški:

- 1 poizkus ropa

Nizko tveganje **3**

Po izpustu: 3 posestva droge



Kriminal: posestvo droge

Predhodni prekrški:

- 1 upiranje pridržanju brez nasilja

Visko tveganje **10**

Po izpustu: /

Nepravična AI je del realnosti

BBC Sign in | Home | News | Sport | Reel

NEWS

Home | Coronavirus | Climate | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts

Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

Tech

Amazon scrapped 'sexist A

10 October 2018

VERDICT

AI AND AUTOMATION

World Diversity Day: Twitter removes AI cropping tool which deleted black people from pictures

Eric Johansson | 21st May 2021 (Last Updated May 21st, 2021 17:07)

By Steve Lohr

Feb. 9, 2018

Science

Current Issue | First release papers | Archive | About | Submit

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAIN

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/sci

	Darker Female	Lighter Male	Lighter Female	Largest Gap
	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

THE WALL STREET JOURNAL.

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr

Updated July 1, 2015 3:41 pm ET

d across the country to predict future criminals.
nd it's biased against blacks.

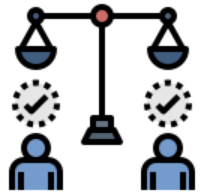
eff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Dve vrsti pravičnosti ML

- **Skupinska pravičnost** – odločanje je nepravično, če velja vsaj eno izmed:
  - **Odločitve** (vsaj delno) **temeljijo** na občutljivih podatkih.
  - **Posledice odločitev nesorazmerno škodujejo** (ali koristijo) ljudem z določenimi vrednotami občutljivih podatkov.
    - Na primer: ženske, starejši, tujci.
  - Proti diskriminacijski zakoni v številnih državah prepovedujejo nepravično obravnavo ljudi na podlagi **občutljivih podatkov**, kot sta spol ali rasa.
    - Zakaj bi pa algoritmi/modeli/računalniki/stroji bili diskriminatorni?
- **Individualna pravičnost** – je odločitev o posamezniku podobna odločitvam podobnih posameznikov?

Občutljivi podatki
(angl. *sensitive/protected data*)



- **Občutljivi podatki** so podatki, na podlagi katerih se ne bi smele delati razlike v odločitvah in v kakovosti odločitev.
  - Največkrat povezani s človeškimi demografskimi podatki, ki so jih kot družba določili: spol, rasa, vera, starost… Včasih je izbira teh kontroverzna in različna med kulturami in zakonodajami.
  - So lahko tudi nečloveški: znamka avtomobila, poreklo naprave, cena izdelka…
  - Niso vedno isti – **so odvisni od problema**. Primer: pri zaposlovanju v splošnem ne želimo, da igra vera pomembno vlogo. Pri zaposlitvi za študentsko delo, pa obstaja omejitev let.

- **Neprivilegirane ali občutljive skupine**, je podmnožica tistih skupin, ki so glede na občutljive podatke v nesorazmerni škodi.
  - Včasih le glede na en občutljivi podatek, včasih v kombinacijami z več teh.

Zakaj pride do skupinske nepravičnosti

- **Pristranosti**
  - **Zgodovinska** (angl. *historical bias)* – model zajeme nepravičen vzorec, ki je prisoten v podatkih zaradi zgodovinskih razmer.
    - Primer: v realnosti so ženske bile manj pogosto zaposlene kot moški, kljub enakim referencam.
  - **Sociološka** (angl. *social bias*) – model zajeme nepravičen vzorec, ki je prisoten v podatkih zaradi trenutnih družbenih/socioloških razmer.
    - Primer: za službo se ne preferirajo tujci.
  - **Reprezentacija** (angl. *representation bias*) – v podatkih so določene senzitivne občutljive skupine v manjšini (ali v drugačnem razmerju kot v realnosti).
    - Primer: večji del naših podatkov FERI študentov je o moških. Kaj pa ženske?

Zakaj pride do skupinske nepravičnosti

- **Ostalo**
  - **Slabi podatki** – niso podatki vseh občutljivih skupin enako kakovostni.
    - Primer: slike iz socialno slabše situiranih skupin (npr. iz Afrike).
  - **Vzorci različnih kompleksnosti** – podatki nekaterih občutljivih skupin imajo težje razberljive vzorce.
    - Primer: temnejše polti se težje analizirajo na slikah.

- TODO: dodaj slike različnih kakovosti.

Posledice nepravične ML

- **Nepravične dodelitve** (angl. *harm of allocation*): Model določenim skupinam ponuja ali odreka priložnosti, vire ali informacije.
  - Primer: Pri zaposlovanju, sprejemu v šolo in dajanju posojil je model veliko boljši pri izbiri dobrih kandidatov med določeno skupino ljudi kot med drugimi skupinami.

- **Nepravična kakovost storitev** (angl. *harm of quality of service)*: Model za eno skupino ljudi ne deluje tako dobro kot za drugo.
  - Primer: Model za prepoznavanje glasu ne deluje tako dobro za ženske kot za moške.

Posledice nepravične ML

- **Nepravično zanemarjanje** (angl. *harm of denigration*): Model določenim skupinam neupravičeno pripiše slabe lastnosti.
  - Primer: Model črnce označi kot gorile.

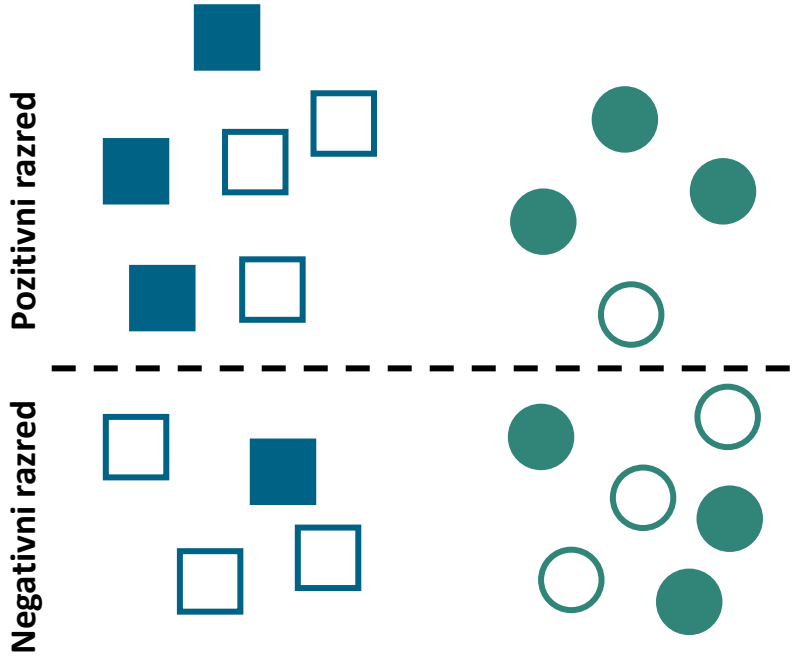
- **Nepravična zastopanost** (angl. *harm of representation)*: Model nadaljuje s prispevanjem k nepravični zastopanosti.
  - Primer: Model še naprej pri zaposlovanju preferira moške napram ženskam.

- **Nepravična stereotipizacija** (angl. *harm of stereotyping*): Model nadaljuje s prispevanjem k nepravični stereotipizaciji.
  - Primer: Ženske ne vozijo avtomobilov, muslimani so teroristi.

Primer v slovenščini

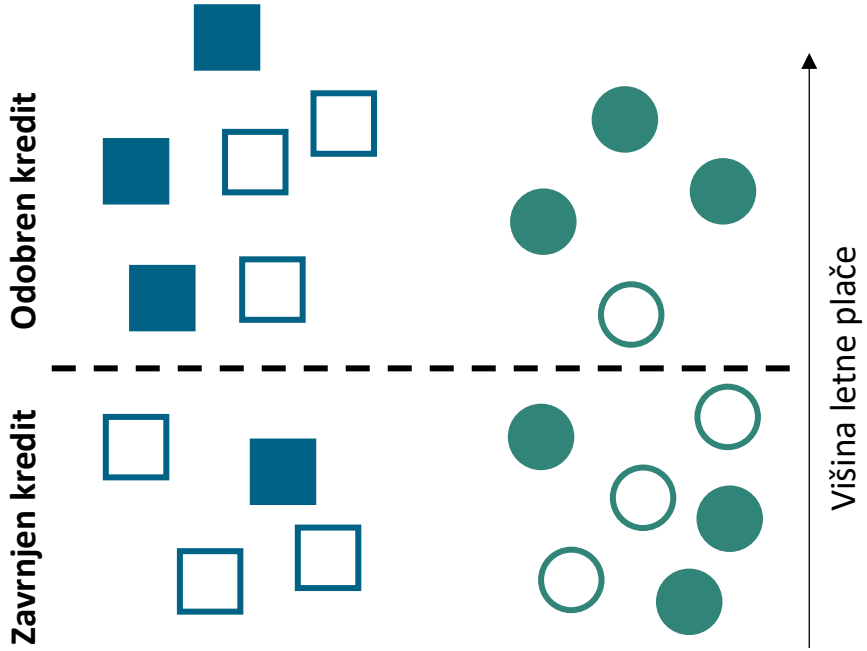
Metrike pravičnosti

Pozitivni razred –
razred, ki je zaželjen
(odločitev, ki je zaželjena)



- Občutljiva skupina 1, zasluženo
- Občutljiva skupina 1, nezasluženo
- Občutljiva skupina 2, zasluženo
- Občutljiva skupina 2, nezasluženo

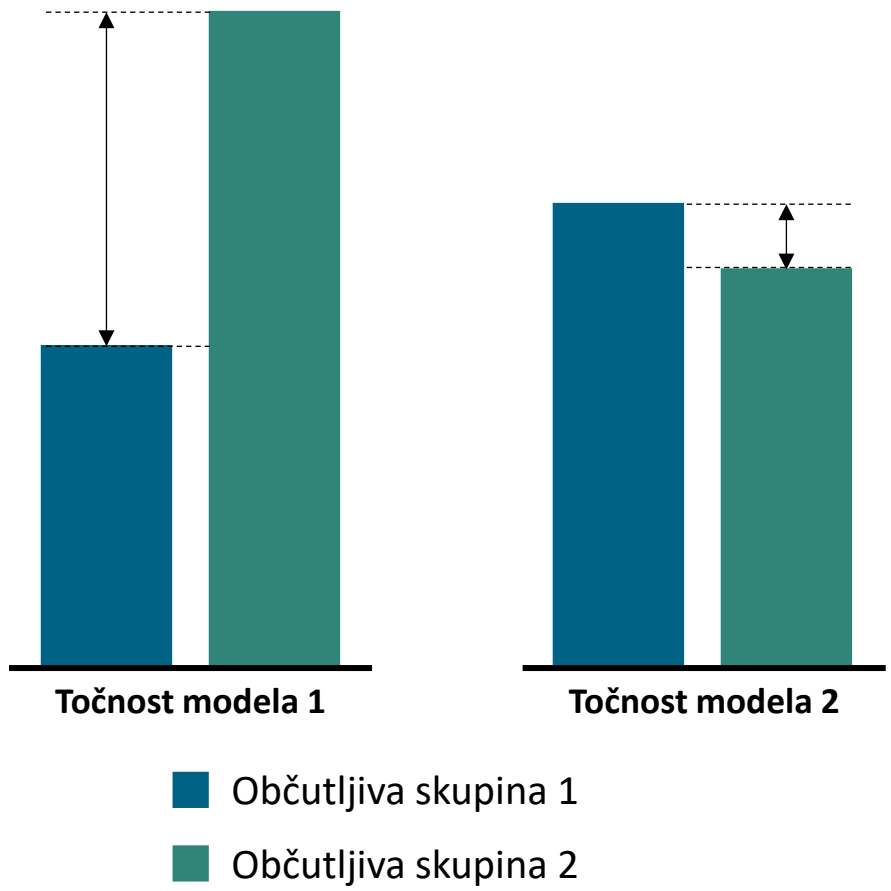
Metrike pravičnosti



- Moški, zaslužili bi si kredit
- Moški, kredita si ne bi zaslužili
- Ženske, zaslužile bi si kredit
- Ženske, kredita si ne bi zaslužile

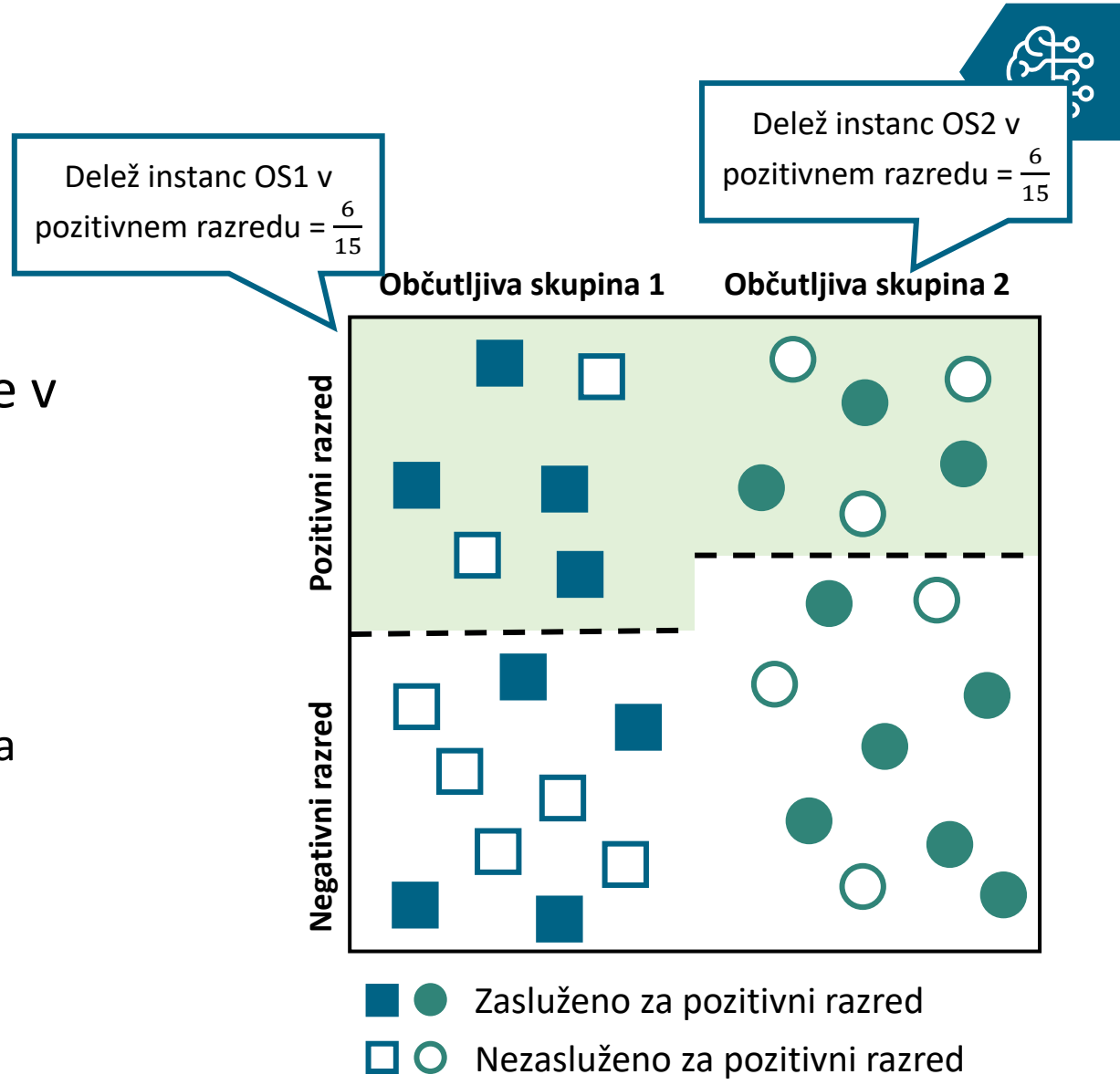
Skupinska pravičnost

- **Enaka kakovost** (angl. *quality equality*) meri, če je kakovost (npr. točnost, F-mera…) napovedi za vse občutljive skupine enaka.
  - Uporabimo: ko želimo, da modeli enako kakovostno odločajo za vse občutljive skupine.
  - Primer iz vsakdana: želimo, da samovozeči avtomobili enako kakovostno prepoznavajo tako otroke, kot odrasle.



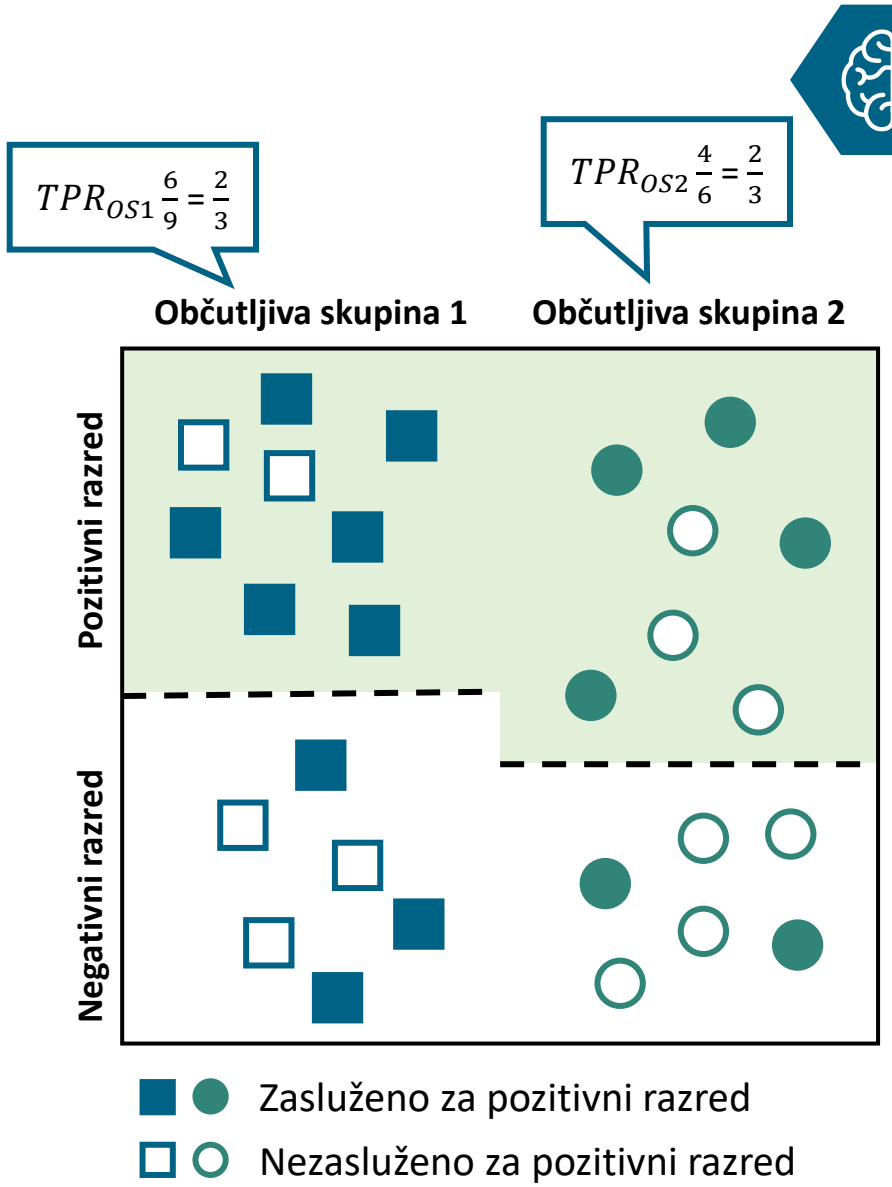
Skupinska pravičnost

- **Demografska enakost** (angl. *demographic* ali *statistical parity*) pomeni enake deleže instanc vseh občutljivih skupin klasificirane v **pozitivni razred**.
  - Uporabimo, ko želim, da je delež instanc, ki so v pozitivnem razredu enak v vseh občutljivih skupinah.
  - Primer: Priporočilni algoritem za priporočanje filmov si prizadeva za demografsko enakost, da bi gledalcem iz različnih starostnih skupin zagotovila enako verjetnost, da bodo videli priporočila različnih žanrov.



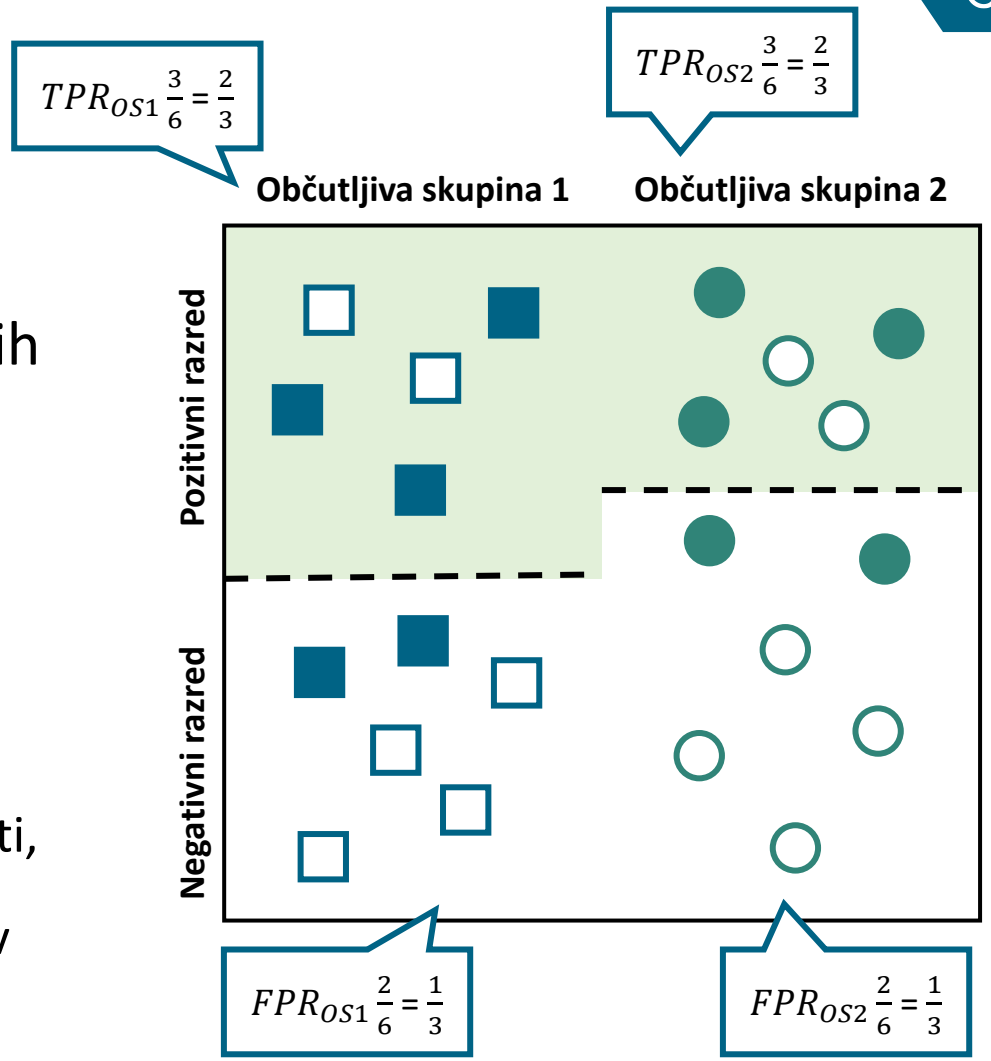
Skupinska pravičnost

- **Enaka priložnost** (angl. *equal opportunity* ali *selection rate*) meri enakost v pravično dodeljenem pozitivnem razredu v vseh občutljivih skupinah.
  - **Pravično dodeljen**: TPR (delež resnično pozitivnih, angl. *true positive rate*) = $\frac{TP}{TP+FN}$
  - Uporabimo, ko želimo zagotoviti, da so deleži med instancami v pozitivnem razredu in tistimi, ki bi si zaslužili, da so v pozitivnem razredu, enaki v vseh občutljivih skupinah.
  - Primer: Tehnološko podjetje v postopku zaposlovanja uporablja algoritem, ki zagotavlja, da imajo kandidati iz vseh okolij (občutljive lastnosti) s potrebnimi znanji (zaslužni) enake možnosti za uvrstitev v ožji izbor.



Skupinska pravičnost

- **Enaka verjetnost** (angl. *equalized odds*) hkrati meri enakost v pravično dodeljenem pozitivnem razredu (kot enaka priložnost) in nepravično dodeljenem pozitivnem razredu v vseh občutljivih skupinah.
  - **Nepravično dodeljen**: FPR (delež lažno pozitivnih, angl. *false positive rate*) = $\frac{FP}{TN+FP}$
  - Uporabimo, ko želimo izenačiti tudi deleže nepravilno dodeljenih pozitivnih razredov.
  - Prime: Finančna ustanova uporablja algoritem za odobritev posojil in si prizadeva za izenačitev možnosti, da bi zagotovila enako odobritev posojil zaslužnim prosilcem in zavrnitev posojil nezaslužnim prosilcem v različnih rasnih skupinah.



Individualna (lokalna) pravičnost

- **Alternativna pravičnost** (angl. *counterfactual fairness*) je, če se NE spremeni odločitev modela, ko se spremeni vrednost občutljive spremenljivke.
  - Uporabno za **eno instanco** (da vidimo, če je za tisto instanco bila narejena odločitev neodvisno od občutljivih lastnosti) ali za **množico instanc** (da merimo delež spremenjenih odločitev).
  - Uporabimo, ko želimo vsiliti neodvisnost napovedi od občutljivih spremenljivk.
  - Primer: Alternativna pravičnost se uporablja v postopku odobritve posojila, pri katerem bi odločitev, ki jo algoritem sprejme za posameznika, ostala enaka, tudi če bi se hipotetično spremenila njegova občutljiva lastnost (npr. rasa ali spol), kar zagotavlja, da izid temelji izključno na ustreznih dejavnikih in ne na diskriminatornih predsodkih.

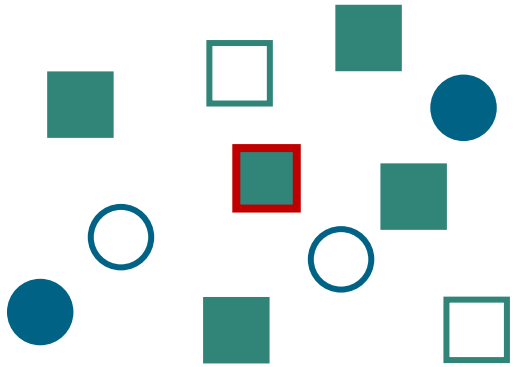
Izkušnje	Znanje	Spol	Odločitev
5	3	M	✓
4	3	Ž	X
5	4	Ž	✓
3	2	M	✓
2	1	M	X

Izkušnje	Znanje	Spol	Odločitev
5	3	Ž	✓
4	3	M	X
5	4	M	X
3	2	Ž	✓
2	1	Ž	✓

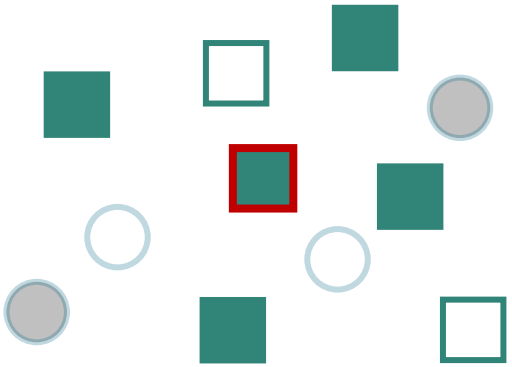
Individualna (lokalna) pravičnost

- **Konsistentnost** (angl. *consistency*) meri, kako konsistentne so bile odločitve za podobne instance.
  - Delovanje: najdemo k najbližjih sosedov izbrani instanci in pregledamo kako podobne so bile odločitve modela.
  - Pri iskanju najbližjih sosedov lahko (a) ignoriramo občutljive spremenljivke, lahko pa (b) vsilimo, da najde najbližje sosede z drugačnimi vrednostmi občutljivih spremenljivk.

Ignoriramo senzitivne spremenljivke



Upoštevamo senzitivne spremenljivke



Katero metrik pravičnosti izbrati?
Na primeru COMPAS

- **Enaka kakovost** pri priprtju zapornikov glede na raso pomeni, da so deleži *pravično* izpuščenih (pozitivni razred) in deleži *pravično* zadržanih (negativni razred) enaki glede na raso.
- **Demografska enakost** pomeni, da bi deleži izpuščenih bili enaki glede na raso.
- **Enaka priložnost** pomeni, da so deleži *pravično* izpuščenih enaki glede na raso.
- **Enaka verjetnost** pomeni, da so deleži *pravično* izpuščenih (pozitivni razred) in deleži *nepravično* izpuščenih enaki glede na raso.
- **Alternativa pravičnost** pomeni, da je odločitev glede izpuščenosti/priprtja za zapornika enaka tudi, če spremenimo raso zapornika.
- **Konsistentnost** pomeni, da je odločitev glede izpuščenosti/priprtja za zapornika enaka tudi v ostalih podobnih primerih (ob upoštevanju ali ignoriranju rase).

Pravičnost iz vidika sodnika.

Pravičnost iz vidika družbe.

Pravičnost iz vidika zapornika (ProPublica).

Pravičnost iz vidika sodnika.

Pravičnost iz vidika zapornika.

Paradoksi pravičnosti

- **Nasprotujoče definicije (in meritve) pravičnosti** – optimiziranje modela za eno vrsto pravičnosti lahko zmanjša drugo vrsto.
  - Ni univerzalno sprejetega standarda pravičnosti.

- **Prava definicija pravičnosti** – katera mera pravičnosti je sploh primerna za dan primer?
  - Različne senzitivne skupine gledajo na pravičnost drugače.

- **Skupinska (globalna) in posamezna (lokalna) pravičnost sta včasih nasprotujoči** – optimiziranje modela za pravičnost za skupine, lahko zmanjša pravičnost na ravni posameznika; ter obratno.

Paradoksi pravičnosti

- **Pravičnost napram splošni kakovosti modela** – poudarek na pravičnosti odločitev največkrat poslabša splošno kakovost napovedi.
  - Je splošna kakovost sploh zaželena?

- **Pravičnost napram zasebnosti** – za merjenje pravičnosti potrebujemo občutljive podatke, ki so pa največkrat zasebni.

- **Nezaželeno učinki pravičnosti** – optimiziranje modela za pravičnost lahko privede do diskriminacije prej nediskriminiranih.

Kako se lotimo nepravičnih AI modelov?

- **Pred procesiranjem** (delo na podatkih)
  - Popravimo razmerja (nad- in pod-vzorčenje) med občutljivimi skupinami.
  - Zbrišemo ali pokvarimo občutljive in občutljivim nadomestne podatke.
- **V procesiranju** (prilagajanje algoritmov za učenje modelov znanja)
  - Prilagodimo notranje metrike algoritmov ML, da poleg splošne kakovosti skrbijo tudi za pravičnost.
- **Po procesiranju** (popravki modelov znanja)
  - Končne modele znanja ovrednotimo glede na pravičnost in jih popravimo, da so bolj pravični.

Tehnike zagotavljanja pravičnosti ML
Pred procesiranjem

- **Pravičnost skozi nevednost** (angl. *fairness through unawareness*) je tehnika, kjer občutljive spremenljivke odstranimo pred procesom učenja modela s ciljem zagotavljanja pravičnosti odločitev.
  - Kot, če bi ljudem prikrili občutljive podatke pred njihovimi odločitvami.
  - Dve stopnji skrivanja:
    - Pred predprocesiranjem podatkov
    - Pred učenjem modela
  - Ne deluje vedno, zaradi nadomestnih podatkov!

Izkušnje	Znanje	Spol	Odločitev
5	3	M	✓
4	3	Ž	X
5	4	Ž	✓
3	2	M	✓
2	1	M	X

Izkušnje	Znanje		Odločitev
5	3		✓
4	3		X
5	4		✓
3	2		✓
2	1		X

Nadomestni podatki
(angl. *proxy data*)

- Včasih so informacije o občutljivih podatkih skrite v drugih “neobčutljivih” podatkih, ki jih mi ne smatramo kot občutljive. Takim pravimo **nadomestni podatki** oz. kvazi-indetifikatorji.

- Primer: če je nekod bil skupaj več kot pol leta na dopustu, je velika verjetnost, da gre za žensko (je šlo za porodniško).

Občutljiva spremenljivka	Primeri nadomestnih spremenljivk
Spol	Izobrazba, dohodek, poklic, podatki o kaznivih dejanjih, ključne besede v besedilu (življenjepis, socialni mediji...), fakulteta, delovni čas
Zakonski stan	Izobrazba, dohodek
Rasa	Podatki o kaznivih dejanjih, ključne besede v besedilu (življenjepis, družbeni mediji...), poštna številka
Invalidnost	Podatki o osebnostnih testih

Tehnike zagotavljanja pravičnosti ML
Pred procesiranjem

- **Pravičnost z alternativami** (angl. *fairness with counterfactuals*) je tehnika, kjer podatkom spremenimo občutljive lastnosti.
  - Kot, če bi se ljudem zlagali glede občutljivih podatkov pred njihovimi odločitvami.
  - S tem pokvarimo vzorce, ki bi naj vsebovali občutljive lastnosti in prisilimo algoritem strojnega učenja, da se nauči odločati neodvisno od občutljivih lastnosti.
  - Več načinov:
    - Delu učne možice naključno spremenimo/dodelimo občutljive lastnosti.
    - Celotni učni možici naključno spremenimo/dodelimo občutljive lastnosti.
    - Nadvzorčimo učno množico z naključno spremenjenimi/dodeljenimi občutljivimi lastnostmi.

Spol	Odločitev
M	✓
Ž	X
Ž	✓
M	✓
M	X

Spol	Odločitev
M	✓
M	X
Ž	✓
M	✓
Ž	X

> *The privileged are processed by people, the poor are processed by algorithms.*

Cathy O’Neil

Avtorica knjige *Weapons of Math Destruction*