



# Neprimerna uporaba strojnega učenja

# Avtomatizacija zaposlovanja na Amazonu



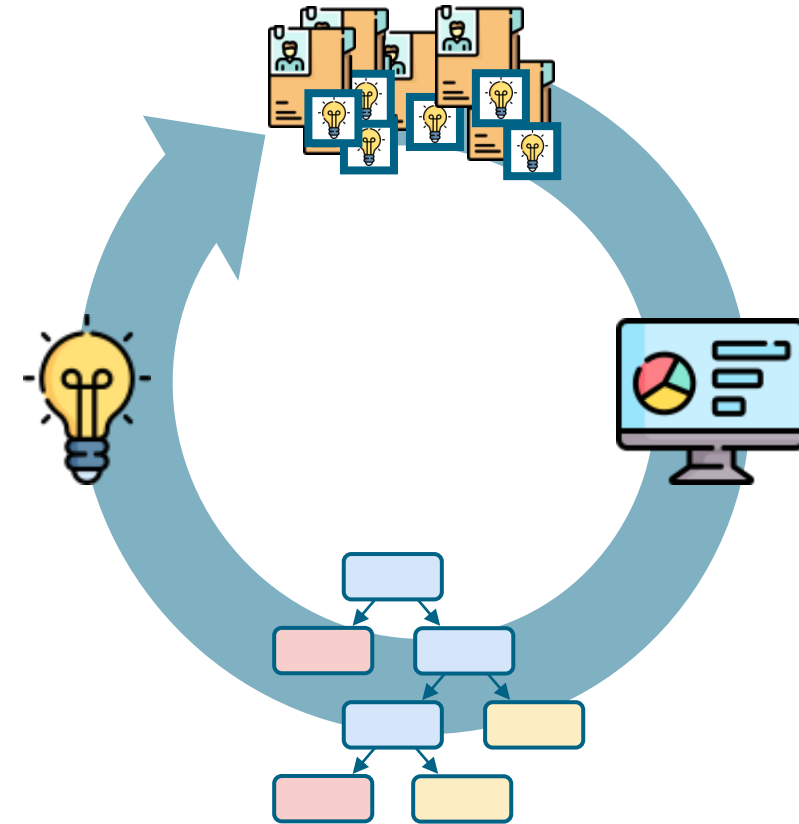
- Pri Amazonu dobijo tisoče prošenj za službo dnevno, ki jih je potrebno pregledati in se odločiti kdo je primeren za udeležbo na interviju za službo in kdo ne.
  - Do sedaj so to delali zaposleni v HR.
  - Kako bi to avtomatizirali?
- Uporabili so spremenljivke iz življenjepisov kot neodvisne in pretekle odločitve zaposlenih HR kot razrede (če je posameznik primeren ali neprimeren za interviju).
- Dobili so seksitičen AI. Zakaj?



# Povratna zanka AI napovedni



- Ljudje domnevajo, da so **algoritmi objektivni in delujejo brez napak**.
  - Zaupajo algoritmom bolj kot ljudem, tudi če jim je na voljo možnost, da ima človek možnost nadzorovati algoritme.
- Izziv: **Podatki**, ki jih uporabljamo neposredno ali posredno **opisujejo odločitve, ki so bile v družbi že sprejete**.
- To ustvari **povratno zanko**, v kateri bo nepoštena odločitev, ki jo bomo sprejeli na podlagi AI, utrjevala nadaljnjo diskriminacijo v družbi in tako še bolj skazila prihodnje podatke.
  - In celo okrepi človeško pristranskost/nepoštenost.



# Upravljanje umetne inteligence



- **Upravljanje umetne inteligence** (angl. *AI alignment*) so pristopi, da sistemi, ki uporabljajo modele znanja naučenih s strojnim učenjem, zanesljivo izvajajo naloge, ki so koristne za ljudi, brez nenamernih negativnih posledic.
  - Je ključnega pomena za varno uporabo prihodnjih, bolj avtonomnih sistemov umetne inteligence.
- Zahteva reševanje **tehničnih izzivov** (kot sta robustnost in posploševanje) in **filozofskih izzivov** (kot je opredelitev človeških vrednot).

# FATE princip



- Namerni ali **nenamerni vplivi AI modelov** na družb postavljajo izzive in pomisleke glede njihove uporabe.
  - Po eni strani pospešujejo tehnološki in družbeni napredek.
  - Po drugi strani, pa hkrati (največkrat za namen večjega zaslužka) vplivajo na posameznike in družbo.
- Z vse pogostejšimi negativnimi vplivi upada zaupanje in volja za vlaganje v AI tehnologijo.
- Izziv: Kako zagotoviti, da bodo ti **vplivi v okviru primernih meja** (ter s tem povečati zaupanje) in hkrati **ne zaustaviti tehnološkega napredka**, ki ga ponujajo.

# FATE princip



- **Fairness (pravičnost)**
  - *So podatki zbrani pravično?*
  - *So odločitve modelov in pravični glede na vse okoliščine (demografijo, čas odločitev, trenutno situacijo)?*
- **Accountability (odgovornost)**
  - *So podatki zbrani in pripravljeni odgovorno (po načelih regulacije)?*
  - *Se modeli zgrajeni odgovorno (po načelih regulacije)?*
- **Transparency (transparentnost)**
  - *So modeli transparentni?*
  - *So odločitve modelov razložljive oz. razumljive?*
- **Ethics (etika)**
  - *Modeli spoštujejo temeljna moralna načela in človekove pravice?*
  - *Vodijo odločitve v nenamerno diskriminacijo ali etične dileme?*

# AI etika (angl. *ethics*)



- Etika na področju AI pomeni, da so sistemi oblikovani in uporabljeni na način, ki **spoštuje temeljna moralna načela in človekove pravice**.
  - Sistemi umetne inteligence ne smejo povzročiti škode posameznikom ali družbi in morajo zagotavljati, da se AI uporablja za splošno dobrobit.
  - AI tehnologije morajo biti oblikovane tako, da spoštujejo etične standarde na vseh stopnjah razvoja in uporabe.
- Razmislimo: diagnoza bolezni, avtomatsko trgovanje z vrednostnimi papirji, orožja (npr. droni)?

# AI odgovornost (angl. *accountability*)



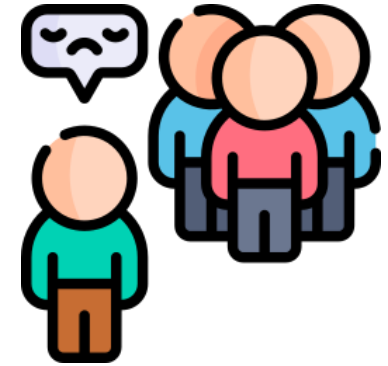
- Odgovornost na področju AI pomeni, da so sistemi, ki uporabljajo AI modele in njihovi ustvarjalci **odgovorni za svoje napovedi in posledice**.
- Princip AI odgovornosti zahteva, da se v primeru napake ali škode, ki jo povzročijo napovedi AI modelov, jasno razume, kdo ali kaj je kriv.
  - Kreatorji AI modelov morajo biti odgovorni za odločitve, sprejete v fazah *načrtovanja*, *učenja* in *integracije* AI modela.
- Princip AI odgovornosti pomaga ohranjati zaupanje javnosti v tehnologije umetne inteligence.
  - Zagotavljajo, da AI spoštuje etične standarde, zakone in predpise.
  - Omogoča deležnikom, da izrazijo dvome glede AI modelov, ne da bi pri tem zavirali inovacije.



# Obstoječa regulativa




- **Proti diskriminacijski zakoni** v številnih državah prepovedujejo nepošteno obravnavanje ljudi na podlagi občutljivih lastnosti, kot sta spol ali rasa.
  - **Listina Evropske unije o temeljnih pravicah** prepoveduje diskriminacijo na podlagi katerega koli razloga, kot so spol, rasa, barva kože, etnično ali socialno poreklo, genetske značilnosti, jezik, vera ali prepričanje, politično ali drugo mnenje, pripadnost narodni manjšini, premoženje, rojstvo, invalidnost, starost ali spolna usmerjenost.
  - **Ameriški zakon o državljanskih pravicah** iz leta 1964 delodajalcem prepoveduje diskriminacijo zaposlenih na podlagi spola, rase, barve kože, narodne pripadnosti in vere.
  - **Japonska ustava** (1946) v 14. členu zagotavlja enakost pred zakonom in prepoveduje diskriminacijo na podlagi političnih, gospodarskih in družbenih razmerij, rase, veroizpovedi, spola, družbenega položaja ali družinskega porekla.
- Če zakon prepoveduje nepoštene odločitve ljudi, kaj pa odločitve, ki jih sprejmejo stroji?



# EU AI Act

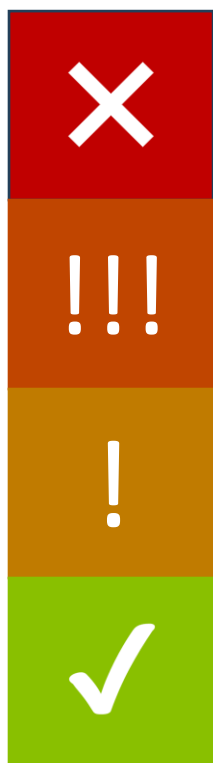


- **EU AI Act** je predlog Evropske unije za urejanje uporabe AI. Namenjen je zagotavljanju, da se AI uporablja na način, ki je varen in spoštuje človekove pravice.
  - Prav tako želi spodbujati inovacije in razvoj AI v Evropi, obenem pa zagotoviti, da AI ne škodi ljudem ali družbi.
  - Predlog prinaša pravila in smernice o tem, kako in kdaj je mogoče uporabljati AI, ter določa kazni za kršitve.
- Zakon ne nadomešča zaščite, ki jo zagotavlja Splošna uredba o varstvu podatkov (GDPR), vendar se z njo prekriva, čeprav je področje uporabe prve obsežnejše in ni omejeno na osebne podatke.
  - V originalni verziji bi že GDPR moral vsebovati dele, predvidenih za AI Act.

# EU AI Act – ravni tveganja



- Zakon bi razvrstil sisteme umetne inteligence glede na tveganje v štiri ravni in zahteval različne ravni ureditve za vsako raven tveganja.



- **Nesprejemljivo tveganje** – prepovedano
  - Socialno ocenjevanje, subliminalne tehnike, realno-časovna biometrična identifikacija v javnih prostorih, izkoriščanje ranljivosti ljudi...
- **Visoko tveganje** – zahtevana revizija
  - Ostala biometrična identifikacija, dostop do socialnih storitev, odločitve o zaposlovanju, napovedovanje kaznivih dejanj...
- **Nizko tveganje** – zahtevana transparentnost
  - Chatboti, biometrično kategoriziranje, uravnavanje čustev, globoki ponaredki (angl. *deep fakes*).
- **Minimalno tveganje** – interna regulacija
  - Vsi drugi sistemi umetne inteligence (spam filtra, igre...).

# EU AI Act – kaj regulira?



- **Uporaba in kakovosti podatkov** – Kako so podatki shranjeni, deljeni in uporabljeni; so podatki reprezentativni, brez napak, pridobljeni zakonito?
- **Transparentnost odločitev** – Ali lahko razumemo in pojasnimo odločitve umetne inteligence?
- **Človeški nadzor** – Kako v proces odločanja vključimo človeka?
- **Odgovornost** – Kdo je odgovoren, če gre kaj narobe?
- Nesprejemljivo tveganje se kaznuje z do 30 milijoni EUR ali 6% svetovnega letnega prometa (kar je višje).
  - Za visoko tvegane sisteme se kaznuje z do 20 milijonov EUR ali 4% prometa.