Od kod izvira tveganje?

Evropska parlamentarna raziskovalna služba je v svoji študiji o umetni inteligenci[1] navedla: "*Važno je opozoriti, da algoritmi umetne inteligence ne morejo biti objektivni, saj tako kot ljudje med usposabljanjem razvijejo način osmišljanja tistega, kar so že videli, in ta 'pogled na svet' uporabijo za kategorizacijo novih situacij, ki so jim predstavljene.*" [deepl translation]

Poglejmo, od kod izvira subjektivnost umetne inteligence in kakšna so s tem povezana tveganja.

Predsodki v podatkih in algoritmih

Tako kot pri vsakem digitalnem sistemu tudi podatki, ki se uporabljajo v platformah, ki temeljijo na umetni inteligenci, prihajajo iz različnih virov in imajo več oblik. V njih so prisotne različne vrste pristranskosti[2]. Predsodki v podatkih so večinoma statistični. Naštejmo nekaj izmed njih.

- **Za vzorčno pristranskost** je značilno, da je prisotna v vrednostih podatkov. To na primer velja za algoritem za zaposlovanje, ki je usposobljen na podatkovni zbirki, v kateri so moški nadpovprečno zastopani, in bo izključil ženske.
- **Stereotipna pristranskost** je nagnjenost k delovanju glede na družbeno skupino, ki ji pripadamo. Ena od študij na primer kaže, da ženske ponavadi kliknejo na ponudbe za delo, za katere menijo, da jih bodo kot ženske lažje dobile.
- **Priznana pristranskost spremenljivke** (pristranskost pri modeliranju ali kodiranju) je pristranskost zaradi težav pri predstavljanju ali kodiranju dejavnika v podatkih. Ker je na primer težko najti dejanska merila za merjenje čustvene inteligence, ta razsežnost ni vključena v algoritme zaposlovanja.
- **Izbirna pristranskost** pa je posledica značilnosti vzorca, izbranega za oblikovanje zaključkov. Banka bo na primer za pridobitev kreditne ocene uporabila interne podatke, pri čemer se bo osredotočila na tiste, ki so ali niso pridobili posojilo, ne bo pa upoštevala tistih, ki nikoli niso potrebovali posojila, itd.

Algoritmična pristranskost je predvsem stvar utemeljevanja. Inženirji umetne inteligence takšno pristranskost uvedejo namerno ali ne.

Prej omenjena študija Evropske parlamentarne raziskovalne službe navaja dva konkretna primera: "*Predpostavimo simbolični algoritem umetne inteligence za pregledovanje prošenj za zaposlitev. Morda bo kandidate ocenil tako, da jim bo dodelil točke samo na podlagi njihove izobrazbe in izkušenj. Če pa algoritem ne upošteva dejavnikov, kot je porodniški dopust, ali*

ustrezno ne prizna izobraževanja v tujih institucijah na način, kot bi to storile človeške izbirne komisije, lahko diskriminira ženske in tuje kandidate." [deepl translation]

"*Tudi si oglejte podobno orodje umetne inteligence v okviru paradigme strojnega učenja (ML). Takšni algoritmi najdejo lastne načine za ugotavljanje, katere vrste kandidatov so bile izbrane v njihovih učnih podatkih. Če v preteklosti pri teh izborih obstajajo strukturne pristranskosti - na primer rasna diskriminacija -, se jih lahko algoritem nauči. Tudi kadar so podatki o narodnosti ali etnični pripadnosti iz podatkov odstranjeni, je ML spreten pri iskanju približkov za osnovne vzorce v drugih podatkih, kot so jeziki, poštne številke ali šole, ki so lahko dobri napovedovalci etnične pripadnosti.*" [deepl translation]

Trije vidiki algoritemskega tveganja

Algoritmično tveganje je mogoče opredeliti na tri načine[3].

- Prvič, gre za **algoritmično omejenost**, ki se lahko nanaša tudi na mnenja, kulturno znanje ali celo poslovne prakse. Algoritmi namreč internetnega uporabnika kljub spoštovanju načela pravičnosti soočajo z enako vsebino, odvisno od njegovega profila in integriranih parametrov. Tako je na spletnih straneh za priporočanje novic, kot je Facebook, ali na spletnih straneh za priporočanje izdelkov, kot je Amazon.
- Drugi vidik algoritemskega tveganja je povezan z **obvladovanjem vseh vidikov posameznikovega življenja**, od urejanja informacij za vlagatelje do njegovih prehranjevalnih navad, hobijev ali celo zdravstvenega stanja. To sledenje posamezniku kaže na obliko nadzora, ki je v nasprotju s samim bistvom svobode posameznika.
- Tretji je povezan z **možnostjo kršitve temeljnih pravic**. Zlasti z algoritemsko diskriminacijo, ki je opredeljena kot neugodno ali neenako obravnavanje v primerjavi z drugimi osebami ali drugimi enakimi ali podobnimi položaji na podlagi razloga, ki je izrecno prepovedan z zakonom. To zajema preučevanje pravičnosti (*pravičnosti*) algoritmov za razvrščanje (razvrščanje ljudi, ki iščejo službo na spletu), priporočanje in učenje napovedovanja. Problem diskriminatorne pristranskosti, ki jo povzročajo algoritmi, zadeva več področij, kot so spletno zaposlovanje, sodne odločitve, odločitve policijskih patrulj ali sprejem v šole.

Kako ravnati s podatkovnimi in algoritmičnimi tveganji?

R. Schwartz in drugi[4] menijo, da "*Pristranskost ni niti nova niti edinstvena za umetno inteligenco in ni mogoče doseči ničelnega tveganja pristranskosti v sistemu umetne inteligence*".
Medtem pa je priznanje, da so agenti UI po naravi subjektivni, ključni predpogoj za zagotovitev, da se uporabljajo le za naloge, za katere so dobro opremljeni.

Študija EPRS se zaključuje z več priporočili pri uporabi aplikacij, ki temeljijo na umetni inteligenci:

- Razumevanje pristranskosti in subjektivnosti
- izogibajte se aplikacijam, ki presegajo zmožnosti umetne inteligence
- Izogibajte se aplikacijam z neželenimi učinki
- Ohranjanje človeške avtonomije
- Iščite rešitve za probleme in ne problemi za rešitve
- Razmislite, kaj si od umetne inteligence resnično želimo.

---

1. Artificial intelligence: How does it work, why does it matter, and what can we do about it ? - Philip Boucher, Scientific Foresight Unit (STOA) - ISBN: 978-92-846-6770-3 - Union Européenne, 2020 ↩
2. Algorithms, Data and Bias: Public Policy Needed, Anne Bouverot, Thierry Delaporte, 2019 ↩
3. Article in French: D'où vient le risque ? Des données et des algorithmes - Serge Abiteboul, Thierry Viéville, 2020 ↩
4. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence - Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, NIST Special Publication 1270 , 2022 ↩