ETIČNE RAZSEŽNOSTI UMETNE INTELIGENCE

Ana Farič in Ivan Bratko

Uvod

Tehnologijo umetne inteligence (UI) sestavljajo razna področja, med njimi metode za splošno reševanje problemov in preiskovanje problemskih prostorov, predstavitev znanja in metode sklepanja, strojno učenje, računalniško razumevanje naravnega jezika, robotika, računalniški vid itd. Relativna pomembnost teh področij se je s časom spreminjala. Od leta 2000 je strojno učenje postalo najpomembnejše in kaže najvidnejši napredek.

Ideje za strojno učenje so se pojavile že zgodaj. Prvi bolje znan praktično uporaben program strojnega učenja je bil ID3 (Quinlan 1979). Kmalu se je pokazalo, da so potenciali strojnega učenja za praktično uporabo izjemni. Dober zgodnji primer je poskus (Bratko in Mulec 1980), ki je verjetno prva uporaba strojnega učenja v medicinski diagnostiki. Zanimivo je, da so nekateri takrat najbolj znani znanstveniki iz UI, posebej ameriški, v obdobju pred 1990 zmotno ocenjevali, da je uporaba strojnega učenja v zahtevnejših domenah preveč optimistična in ni realna. S tem so močno upočasnili uveljavljanje uporabe strojnega učenja po svetu. Tako je šele okrog leta 2000 strojno učenje postalo najperspektivnejše področje UI.

Vendar je popoln preobrat v vsesplošnem zanimanju za strojno učenje sledil šele po dodatnem tehničnem preboju globokega učenja z nevronskimi mrežami (LeCun, Bengio in Hinton 2015) in generativne UI po letu 2020 (npr. Bubeck et al. 2023). Generativna UI je posebna vrsta strojnega učenja. Pri klasičnem strojnem učenju gre tipično za učenje napovedovanja »razreda«, npr. v medicini za klasifikacijo novega pacienta v diagnostične razrede ali napovedovanje temperature v vremenski napovedi. Pri generativni umetni inteligenci pa se sistem nauči generirati nove *vsebine*, npr. besedila ali slike. Tako lahko na primer v medicinski aplikaciji generira splošen besedilni opis razlik in podobnosti med dvema boleznima.

Ta razvoj je v kratkem času omogočil ogromen napredek v vrsti aplikacij UI. Znani primeri so: kakovostno strojno prevajanje, uporaba v zdravstvu, znanosti (npr. pisanje znanstvenih člankov, oz. vsaj velika pomoč pri tem), izobraževanju, sijajni rezultati, ki jih dosegajo veliki jezikovni modeli na specialističnih izpitih iz medicine in prava, avtomatsko programiranje in odlični rezultati na tekmovanjih iz programiranja.

Vendar so se ob teh tehnoloških dosežkih v UI, še posebej ob pojavitvah velikih jezikovnih modelov, kot je ChatGPT, pojavila tudi relevantna vprašanja o tem, kje so meje zaupanja v to tehnologijo. Ob pojavitvi sistema GPT-4 leta 2023 je prišlo do velikega miselnega preobrata med nekaterimi najvidnejšimi strokovnjaki za UI. Ob uspehih generativne UI so še pred pojavitvijo GPT-4 ti eksperti pogosto komentirali

presenetljive uspehe npr. GPT-2 tako: »Da, zanimivo, toda ti sistemi nimajo pojma, o čem govorijo.« Kljub njihovi presenetljivi zmožnosti je uporabnik takratne jezikovne modele z lahkoto z nekaj vprašanji spravil v težave, v katerih so gladko odpovedali. Toda ko se je pojavil GPT-4, so isti eksperti zelo hitro spremenili mnenje iz »ne vedo, kaj govorijo«, v smeri: »Da, GPT-4 deluje presenetljivo dobro, tega nismo mogli pričakovati.« Vendar ni jasno, *zakaj* deluje tako dobro, saj nihče ne ve dobro, kaj se v resnici dogaja v tej ogromni nevronski mreži s 1000 milijardami parametrov. Komentarji ekspertov so se hipoma spremenili. Zdaj ne vemo več, ali »ne ve, kaj dela«, ali mogoče vendarle »ve, kaj dela«. Tudi eden od najpomembnejših razvijalcev globokega učenja, Geoffrey Hinton, ki je leta 2024 za dosežke na tem področju prejel Nobelovo nagrado, deli zaskrbljenost glede dejstva, da tej tehnologiji zaupamo tudi izjemno pomembne aplikacije, obenem pa ne razumemo, na kakšen način doseže tako prepričljivo delovanje. A obenem dela tudi očitne napake, ki pa jih je zaradi našega nerazumevanja nemogoče predvideti.

Zato so ta razvoj spremljala vprašanja, ali je ta tehnologija varna in »vredna zaupanja« (»*trustworthy AI*«), kakšne so posledice za delovanje družbe. Nekateri razlogi za skrb so: nerazumljivost nekaterih sistemov UI, možnosti manipulacije z ljudmi z metodami UI in s tem tudi vpliv na demokratične procese, vpliv na izobraževanje in tudi na cilje izobraževanja, večanje neenakosti, diskriminacija določenih skupin, npr. glede na raso ali spol.

Hiter razvoj in eksplozija novih načinov uporabe UI močno vplivata na naš način dela in življenja. Zato so se v tej zvezi vse pogosteje pojavljala vprašanja o t. i. etičnih vidikih UI. Te vidike je treba nasloviti, če želimo zagotoviti UI, ki bo vredna zaupanja, njena uporaba pa varna in etična.

V tem poglavju naredimo pregled etičnih vidikov UI. Nato se osredinimo na podrobnejšo razpravo o dveh od najpogosteje omenjanih etičnih vidikov: (1) pristranskost UI in strojnega učenja in (2) transparentnost (razumljivost, razložljivost) sistemov UI ter sposobnost razlage.

Etični vidiki umetne inteligence

Razvoj umetne inteligence so spremljala tudi ugibanja o možnih nevarnostih, ki bi se lahko pojavile ob razvoju te tehnologije. Zgodnje razmišljanje se je ukvarjalo z vprašanjem, ali lahko UI prevzame oblast. Razmišljanje je šlo npr. v smeri inteligentnih robotov, ki naj bi si nasilno podredili ljudi. To so bile v glavnem naivne predstave, ki niso imele dobre osnove v realnosti. Tak scenarij ni bil nikoli videti verjeten, saj bi morebitno nevarnost nasilnih robotov ljudje z lahkoto zaznali in preprečili.

Druga vrsta vprašanj je bila povezana s predvidevanji o t. i. točki tehnološke singularnosti. To je trenutek, ko bi se pojavila superinteligenca, to je umetna inteligenca, ki bi presegla človeško (Kurzweil 2005). Predvidevanja o tem, kdaj bi naj se to zgodilo, so bila zelo nejasna že zato, ker v teh razpravah ni bilo dobro definirano, kdaj bomo šteli, da je UI presegla človeško inteligenco. Inteligenco sestavlja vrsta sposobnosti (npr. pomnjenje, računanje, logično sklepanje, učenje, razumevanje in sporazumevanje v naravnem jeziku, sklepanje »po zdravi pameti«, kreativnost …). V nekaterih od teh komponent inteligence je UI že zdavnaj presegla sposobnosti človeške

ka, v nekaterih pa je bila vsaj do nedavna videti še zelo daleč. Tudi ni bilo jasno, kaj bi bile konkretne nevarnosti, povezane s točko singularnosti. Zato med raziskovalci UI scenarij s singularnostjo ni bil deležen velike pozornosti. Bostromovo (2014) razmišljanje o možnih nevarnostih superinteligence je prepričljivejše. Gre v smeri t. i. eksistenčne grožnje UI za obstoj človeštva.

Konkretne podrobnosti možnih nevarnosti, povezanih z UI, so bile dolgo zelo špekulativne. Zadeva se je spremenila, ko je UI dosegla velik praktičen napredek. Takrat so postali nekateri vidiki možnih nevarnosti konkretnejši in verjetnejši. Leta 2015 je skupina raziskovalcev UI opozorila na te vidike z odprtim pismom (dostopno na spletni strani Future of Life Institute 2015), ki ga je podpisalo več kot enajst tisoč strokovnjakov za UI. Glavno sporočilo pisma je bilo: UI je dosegla velik tehnični napredek, ki omogoča velike priložnosti, obenem pa je prišel čas, da smo pozorni in se izognemo možnim nevarnostim.

Besedna zveza »etični vidiki umetne inteligence« pomeni vprašanja, principe in metode v zvezi z zagotavljanjem varne in etične uporabe UI. Z etično uporabo mislimo na prakso uporabe UI, ki ni v nasprotju s splošno sprejetimi vrednotami in načeli. Med temi se v okviru etične uporabe UI najpogosteje omenjajo naslednje lastnosti: pravičnost in nepristranskost (npr. glede na raso ali spol), transparentnost in razumljivost, spoštovanje zasebnosti (kot nasprotje nelegitimnemu zbiranju osebnih podatkov in njihovi zlorabi), korektnost (verodostojnost) informacij kot nasprotje lažnim novicam (»*fake news*«) in zlonamerni manipulaciji z ljudmi (npr. v politične namene). Angleške fraze, ki pomenijo bolj ali manj isto kot »*etični vidiki UI*«, so: »*Ethical issues of AI, Ethics of AI, AI ethics*«. S temi principi se ukvarja med drugimi tudi mednarodna organizacija Int. Association for Safe and Ethical AI (IASEAI), ustanovljena leta 2025.

V nadaljevanju tega razdelka so predstavljeni etični vidiki UI, ki so bili omenjeni v odprtem pismu iz leta 2015 ali so se izkristalizirali nekoliko pozneje. V naslednjih dveh razdelkih pa so nekateri od teh vidikov obdelani podrobneje.

Transparentnost in razložljivost sistemov UI, razlaga odločitev, problem črnih škatel

Načelo transparentnosti in razložljivosti pomeni, da naj bi imel uporabnik vpogled v delovanje sistema UI in da je sistem zmožen uporabniku razložiti svoje odločitve v človeku razumljivi obliki. Ta princip je posebej bistven, kadar so predlagane odločitve umetne inteligence za uporabnika pomembne, še posebej takrat, kadar obstaja možnost, da je sistem predlagal napačno odločitev (npr. v medicinski diagnostiki).

Nekatere metode UI zlahka, že po osnovnem načinu delovanja, zadostijo načelu razložljivosti. Primer je metoda za učenje odločitvenih dreves. Za odločitvena drevesa se navadno smatra, da so sama po sebi dobro razumljiva. Primer slabe razložljivosti pa je učenje z nevronskimi mrežami, še posebej globokimi mrežam, ki so nerazumljive že zaradi velikosti, navadno merjene s številom prilagodljivih parametrov v mreži. Odločitev nevronske mreže je rezultat skupnega delovanja velikega števila parametrov in je zelo težko, pogosto praktično nemogoče razložiti, kako so številne interakcije v mreži privedle do končne odločitve. Pri tem omenimo, da prav globoke nevronske mreže pogosto dosegajo največjo klasifikacijsko točnost. Zato včasih pride

do neke vrste tehtanja med točnostjo in razložljivostjo, ko v strojnem učenju lahko dosežemo boljšo razložljivost, vendar na račun točnosti.

Zaradi potrebe po razložljivosti je nastalo posebno področje UI, imenovano razložljiva UI (»*explainable AI*«, krajšano kot XAI). To področje se ukvarja z metodami, kako avtomatsko generirati človeku prijazno razlago odločitev sistema UI. Poseben izziv je, kako generirati razlago odločitev v primerih, ko je postopek računanja odločitev zaradi kompleksnosti nerazložljiv. Kot bo pojasnjeno v nadaljevanju, takrat pridejo v poštev metode, ki privedejo do enake odločitve, vendar na preprostejši način, ki omogoča vsaj približno razlago.

Pristranskost v strojnem učenju

Nekateri modeli UI, dobljeni kot rezultat strojnega učenja, so se v praksi izkazali kot pristranski, tipično glede na raso ali spol. Npr. sistem COMPAS (Angwin et al. 2016), ki se uporablja v ameriškem pravosodju, vzbuja vtis, da je diskriminatoren do temnopoltih obtožencev. Za nekatere druge odločitvene sisteme je bilo ugotovljeno, da so diskriminatorni do žensk, npr. v zaposlovanju. To temo so preučevali v raznih raziskavah. Te kažejo zapletenost teh vprašanj. Razni smiselni formalni kriteriji za ugotavljanje pristranskosti so med seboj nezdružljivi. Odločanje med možnimi načini za preprečevanje diskriminacije pa je pogosto stvar zapletene presoje. V razdelku o pristranskosti v nadaljevanju podrobneje predstavimo nekatere ugotovitve in smeri teh diskusij.

Manipulacija z ljudmi z orodji UI

Metode UI je možno učinkovito zlorabiti za manipulacijo z ljudmi po družbenih medijih. Namen take manipulacije je lahko komercialen ali političen, pri čemer lahko na ta način pride do vplivanja na demokratične procese – manipulacije volitev in referendumov. Orodja UI omogočajo avtomatsko generiranje osebno prilagojenih tendencioznih vsebin ali dezinformacij in ciljano širjenje sporočil na tak način, da so doseženi čim večji učinki in spremembe mnenja pri ciljanih posameznikih. Poglavje v tej knjigi Košmrlj in Bratko (2025) se podrobneje ukvarja z vplivom generativne UI na demokratične procese.

Zbiranje osebnih podatkov na spletu in njihova zloraba

Avtomatizirana manipulacija z ljudmi po družbenih medijih in spletu je učinkovita ob uporabi osebnih podatkov o posameznikih ter njihovih lastnostih in nagnjenjih. Zato ponudniki spletnih storitev nenehno zbirajo podatke o uporabnikih, tudi osebne, in jih uporabljajo za ciljano oglaševanje in širjenje tendencioznih vsebin ter si jih prodajajo med seboj. Evropska zakonodaja (GDPR) tako zbiranje podatkov sicer omejuje in ga dovoljuje ob soglasju uporabnika. Vendar zbiralci podatkov tako soglasje uporabnika praktično izsilijo, ne da bi bilo uporabniku res jasno, kako bodo podatki uporabljeni.

Globoki ponaredki

Na tehnologiji globokih nevronskih mrež temeljijo tudi programska orodja za generiranje globokih ponaredkov (»*deep fake*«) ponarejenih video in avdio vsebin.

Zaradi sposobnosti za neprečljivo ponarejanje slike ali govora konkretnih znanih oseb so ti ponaredki praktično nerazpoznavni in omogočajo učinkovito širjenje neresnic po družbenih medijih. S tem dodatno povzročajo splošno informacijsko negotovost.

UI v avtonomnih orožjih

Avtonomna orožja samostojno, brez posredovanja človeškega operaterja, izbirajo tarčo in se odločijo za napad. Pri tem je lahko odločitev napačna in usodna za nedolžne žrtve. Nekateri si razlagajo, da so s tem, ko je odločitev za napad sprejela UI, človeški operaterji razbremenjeni za odgovornost v primeru napake in nedolžnih žrtev. S tem se ob uporabi avtonomnih orožij standardi odgovornosti zmanjšujejo in mnogi, vključno s številnimi strokovnjaki UI, zato menijo, da z uvajanjem avtonomnih orožij prihaja do še ene oboroževalne tekme. Prav zaradi domnevne manjše odgovornosti ljudi (in prenašanja odgovornosti za nedolžne žrtve na umetno inteligenco) se zmanjšuje prag za odločitve držav za vojaške konflikte in posege. V tej smeri gredo tudi pričakovanja države agresorke, da bo uporaba avtonomnih orožij zmanjšala število lastnih žrtev vojne. S tem lahko odločitev za vstop v vojno postane preveč lahkotna. Ta bojazen je tudi glavno sporočilo odprtega pisma, ki je nastalo med konferenco IJCAI 2015 (Int. Joint Conf. on AI, tj. najprestižnejša klasična vsakoletna konferenca iz UI). To pismo je zbralo več kot 34 tisoč podpisov (Future of Life Institute 2015).

Avtomatsko razpoznavanje čustev

Strojno učenje je omogočilo tudi učenje razpoznavanja človekovih čustev iz izraznih vzorcev na obrazu, ki se pojavljajo pri osebah ob določenih čustvenih stanjih. Ti vzorci omogočajo tudi razpoznavanje čustev, ki jih sicer oseba ne bi rada izdala. Posebej avtomatsko razpoznavanje prikritih čustev je etično sporno, saj naj bi imel človek pravico, da sam odloča o tem, ali čustva pokaže ali ne.

Pristranskost v umetni inteligenci[1]

Pojav pristranskosti v umetni inteligenci

Z naraščajočo uporabo strojnega učenja so se v zadnjih 5–10 letih pojavili primeri aplikacij, deležni izrazito odklonilnih odzivov tako s strani splošnih medijev kot znotraj strokovne literature. Pogosto so izpostavljeni sistemi, uporabljeni v domenah pravosodja, zaposlovanja in bančništva. Kritiki opozarjajo, da so algoritmi in sistemi strojnega učenja nepravični in pristranski glede na t. i. zaščitene atribute, kot so rasa, spol in starost posameznika. Trdijo, da odločitve umetne inteligence (UI) temeljijo na teh atributih namesto na objektivni oceni dejstev (Hellström, Dignum in Bensch 2020). Strokovnjaki z različnih področij obravnavajo t. i. problem pristranskosti strojnega učenja. Skušajo opredeliti, kaj pristranskost pomeni, od kod naj bi izhajala in, najpomembneje, kako jo ustrezno nasloviti.

Na razvijajočem se področju etike v UI (npr. UNESCO 2021) se vprašanje pristranskosti strojnega učenja uvršča med osrednje teme. Politiki ga pogosto omenjajo v povezavi z regulativnimi načeli, namenjenimi zagotavljanju etične in varne

1 Vsebina razdelka o pristranskosti UI je bila deloma objavljena v Farič in Bratko (2023).

uporabe UI, npr. Artificial Intelligence Act (Evropski parlament 2024). Vendar pa v teh razpravah pogosto ni jasno opredeljeno, kaj pristranskost strojnega učenja in UI pravzaprav pomeni. Posledično regulativni ukrepi na tem področju ostajajo pomanjkljivo definirani in zgolj na zelo abstraktni ravni. Izraz pristranskost v kontekstu strojnega učenja za različne avtorje pomeni različno. Celo v strokovni literaturi s področja UI ni popolnega soglasja in splošno sprejete tehnične definicije pristranskosti, ki bi omogočala operativno uporabo za njeno preprečevanje (Hellström, Dignum in Bensch 2020). Poleg tega je za razne smiselne definicije pristranskosti matematično dokazano, da jim, razen v posebej trivialnih primerih, ni mogoče zadostiti hkrati (Hüllermeier, Fober in Mernberger 2013).

V nadaljevanju pregledamo različne definicije pristranskosti in različna mnenja o tem, kako naj bi se tega problema v praksi najuspešneje lotili. Različna mnenja najprej ilustriramo z znanim kontroverznim primerom, s sistemom COMPAS. Zaključki nakazujejo, da je za ustrezno obravnavo nujno upoštevati družbene vrednote in jih z interdisciplinarnim sodelovanjem operacionalizirati z demokratično sprejetim družbenim dogovorom v obliki ustrezne zakonodaje. K boljšemu splošnemu razumevanju pristranskosti v UI v praksi bi prispevalo tudi izboljšanje splošne izobrazbe o UI in njenih metodah.

Primer: sistem COMPAS

Neenotnost na področju obravnave pristranskosti bomo v nadaljevanju pokazali na primeru sistema COMPAS (Angwin et al. 2016; Dressel in Farid 2018; Flores, Bechtel in Lowenkamp 2016; Holsinger et al. 2018). Sistem COMPAS je v številnih publikacijah obravnavan kot verjetno najbolj kontroverzen primer, ki ponazarja domnevno pristranskost delovanja UI. COMPAS (»*Correctional Offender Management Profiling for Alternative Sanctions*«) je odločitveni sistem, ki ga na mnogih ameriških sodiščih uporabljajo sodniki za oceno tveganja povratništva. Sistem ocenjuje verjetnost, da bo obsojenec v dveh letih po izpustu ponovno storil kaznivo dejanje. COMPAS je razvilo ameriško podjetje, takrat imenovano Northpointe (danes Equivant). Sistem upošteva 137 atributov o posameznem prestopniku. Podatki so pridobljeni bodisi od obsojenca bodisi iz njegove kriminalne datoteke. Te podatke nato analizira poseben algoritem, ki kot poslovna skrivnost podjetja ni splošno znan. Na podlagi analize algoritem poda oceno med 1 in 10, pri čemer višja ocena pomeni višje tveganje za povratništvo.

Razprava o sistemu COMPAS se je začela z odmevnim člankom v ProPublici (Angwin et al. 2016). Skupina preiskovalnih novinarjev je predstavila svojo analizo sistema COMPAS in eksperimentiranje z realnimi podatki o več kot 7000 obtožencih na Floridi v letih 2012 in 2013. Ugotovili so, da je samo 61 odstotkov ljudi, ocenjenih kot verjetni povratniki, dejansko ponovilo kaznivo dejanje. V nadaljnji analizi so se osredinili predvsem na rasni vidik in prišli do zaključka, da je sistem pristranski do temnopoltih obtožencev. Spremljali so, koliko obtožencev, ocenjenih z visokim tveganjem za povratništvo, je bilo v naslednjih dveh letih dejansko ponovno obsojenih. Zanimalo jih je, kako pogosto so se napovedi sistema COMPAS razlikovale od dejanskih izidov. Posebej sta zanimivi dve vrsti napak: (1) napačen pozitivni primer, ko sistem napačno napove ponovitev prestopka, in (2) napačen negativni primer, ko sistem napačno napove, da obtoženec ne bo ponovil prestopka. Tovrstne napake v strojnem učenju

običajno merimo z meriloma FPR in FNR, ki sta definirani kot: (1) FPR (»*false positive rate*«) je delež napačno klasificiranih primerov med primeri, ki so ocenjeni kot pozitivni, in (2) FNR (»*false negative rate*«) je delež napačnih klasifikacij med tistimi primeri, ki so ocenjeni kot negativni. Zanimalo jih je, ali sta ti dve meri različni za belopolte in temnopolte obtožence. Dobljeni rezultati so v naslednji tabeli. Rezultati nakazujejo, da je COMPAS prijaznejši do belopoltih obsojencev:

	Belopolti	Temnopolti
FPR	23,5 %	44,9 %
FNR	47,7 %	28,0 %

Te rezultate so Angwin et al. (2016) interpretirali kot očitno pristranske do temnopoltih obsojencev in zato ocenili uporabo sistema COMPAS kot neprimerno in diskriminatorno. Taki interpretaciji bi bilo težko nasprotovati.

Dodaten problem, ki ga izpostavijo Angwin et al. (2016), je pomanjkanje transparentnosti pri odločanju sistema COMPAS, saj je algoritem zaščiten kot poslovna skrivnost. Poleg tega sistem COMPAS ne ponudi razlage za svoje napovedi. Zaradi pogoste citiranosti tega članka je COMPAS postal najbolj znan primer pristranskosti v strojnem učenju – tako v strokovnih krogih na področju strojnega učenja kot tudi med širšo javnostjo brez strokovnega znanja o UI. Kljub temu se COMPAS še vedno uporablja.

Na članek v ProPublici (Angwin et al. 2016) je odgovorila skupina strokovnjakov iz ameriškega pravosodja z delom z zgovornim naslovom »False positives, false negatives and false analysis: A rejoinder to Machine Bias …« (Flores, Bechtel in Lowenkamp 2016). V njem so opozorili na več spornih odločitev v analizi ProPublice, izvedli lastno eksperimentalno raziskavo in zaključili, da so trditve v ProPublici (Angwin et al. 2016) napačne. Čeprav se ta kritika zdi upravičena, bi bila bolj prepričljiva, če bi avtorji jasno pokazali, kje točno je prišlo do ključne napake v analizi ProPublice. Namesto tega so predstavili svoj eksperimentalni rezultat, ki naj bi dokazal, da so osumljenci obravnavani enakopravno, ne glede na raso. Do tega rezultata so prišli tako, da so analizirali ocene tveganja obsojencev na lestvici od 1 do 10, kot jih oceni COMPAS. Iz teh ocen so izračunali mero AUC (površina pod ROC-krivuljo), ki se v strojnem učenju tudi pogosto uporablja kot indikator uspešnosti modela. Mera AUC je zanimiva zato, ker predstavlja verjetnost, da napovedni sistem pravilno loči med pozitivnimi in negativnimi primeri. To pomeni, da če vzamemo naključna primera (obtoženca, izmed katerih je prvi ponovil kaznivo dejanje, drugi pa ne), bo sistem z verjetnostjo, enako vrednosti AUC, pravilno določil, kateri je pozitiven in kateri negativen. Po navedbah Floresa, Bechtela in Lowenkampa (2016) je vrednost AUC za belopolte osebe znašala 0,69, za temnopolte pa 0,70. Razlika med vrednostma ni statistično značilna. Iz tega so sklepali, da COMPAS očitno ni rasno diskriminatoren in da rezultati ProPublice, ki kažejo na diskriminacijo, ne morejo biti pravilni. Vendar tak posredni argument dopušča dvom, saj mere AUC ter FPR in FNR med seboj niso enoznačno povezane.

Dressel in Farid (2018) v svoji raziskavi poročata o relevantnem poskusu, v katerem sta raziskovala napovedno točnost naključno izbranih posameznikov brez

domenskega znanja in točnost preprostega linearnega klasifikatorja. Primerjala sta napovedno točnost sistema COMPAS z uspešnostjo omenjenega linearnega klasifikatorja. Eksperiment z napovedovanjem človeških udeležencev (izveden z množičenjem, »crowd sourcing«) sta izvedla na podmnožici podatkov s Floride, ki je zajemala 1000 od skupno 7000 obtožencev iz študij (Angwin et al. 2016; Flores, Bechtel in Lowenkamp 2016). Ker bi bila uporaba vseh 137 atributov iz izvorne zbirke podatkov za poskus s človeškimi udeleženci nepraktična, sta izbrala zgolj sedem atributov. Ti niso vključevali rase. Presenetljivo je, da je bila napovedna točnost neekspertov v teh poskusih praktično enaka kot napovedna točnost sistema COMPAS. Zanimivo je tudi, da so bile človeške napovedi v tem poskusu podobno pristranske kot napovedi sistema COMPAS, merjeno z vrednostmi FPR in FNR za belopolte in temnopolte. Poleg tega so rezultati ostali skoraj nespremenjeni, tudi če so človeški ocenjevalci prejeli dodatne informacije o rasi. Avtorja sta prav tako ugotovila, da preprost linearni klasifikator doseže podobno napovedno točnost, pri čemer uporablja zgolj dva atributa. Poleg tega bolj sofisticirani klasifikatorji niso izboljšali niti napovedne točnosti niti pravičnosti.

V članku iz pravosodja so Holsinger et al. (2018) kritizirali raziskavo Dresselove in Farida (2018), pri čemer so kritiko utemeljili na naslednjih argumentih. Udeleženci v raziskavi (Angwin et al. 2016) so bili rekrutirani na platformi Amazon Mechanical Turk, kjer so za sodelovanje prejeli finančno nagrado. Oceno povratništva so podajali na osnovi sedmih atributov: starost, spol, vrsta kaznivega dejanja, resnost kaznivega dejanja, število obsodb v odrasli dobi, število obtožb za mladoletniška kazniva dejanja in število obtožb za mladoletniške prekrške. Vsi omenjeni atributi so znani kot pomembni dejavniki tveganja za povratništvo. Po mnenju avtorjev je zmanjšanje števila atributov pomembno olajšalo nalogo napovedovanja v primerjavi z izvorno nalogo, ki je vključevala vseh 137 atributov. Taka kritika je utemeljena, saj je znano, da je izbira ustreznih atributov v strojnem učenju lahko izjemno zahtevna. Poleg tega so udeleženci po podaji posamezne ocene prejeli povratno informacijo o pravilnosti odgovora in svoji povprečni natančnosti. Tisti, ki so dosegli višjo stopnjo točnosti, so bili nagrajeni z nekoliko višjim plačilom. Tak eksperimentalni okvir pa se bistveno razlikuje od resničnega konteksta, pri čemer se odločevalci soočajo s kopico (pogosto nerelevantnih in pristranskih) informacij. Glede na vse te okoliščine so avtorji zaključili, da raziskava Dressela in Farida (2018) ni prispevala pomembnih novih spoznanj. Pokazala je zgolj, da lahko naključno izbrani ljudje na podlagi izbranih relevantnih dejavnikov in povratnih informacij bolje napovedujejo povratništvo kot strokovnjaki, ki delujejo v kompleksnem okolju resničnih primerov. Holsinger et al. (2018) pa niso pojasnili, zakaj si tudi strokovnjaki ne bi pomagali z že obstoječim znanjem o pomembnih izbranih atributih. Tako bi si tudi strokovnjaki olajšali nalogo s poenostavitvijo kompleksne informacije v majhno število relevantnih dejavnikov.

Neodvisno od teh rezultatov je Cynthia Rudin (2019) na podlagi strojnega učenja sintetizirala preprost in popolnoma razumljiv napovedni model, temelječ na podatkih s Floride (Angwin et al. 2016). Model sestoji iz pravil »if-then« (navedenih v nadaljevanju) in uporablja zgolj tri atribute: spol, starost in število prejšnjih kaznivih dejanj. V primerjavi s sistemom COMPAS so ta pravila trivialno razumljiva:

```
IF age between 18-20 and sex is male
  THEN predict arrest (within 2 years)
  ELSE IF age between 21-23 and 2-3 prior offenses
    THEN predict arrest
    ELSE IF more than three priors
        THEN predict arrest
        ELSE predict no arrest
```

Ta napovedni model dosega praktično enako točnost kot COMPAS. Ima pa tudi podobne mere FNR in FPR (po podatkih s Floride). Glede na to je tudi ta model, ki ga je iz učnih primerov generiral nedvomno nepristranski algoritem učenja, pristranski v enakem pogledu kot COMPAS. Podobne rezultate za FPR in FNR sta dobila tudi Dressel in Farid (2018) za učenje iz istih podatkov z enostavno linearno regresijo. Zanimivo je, da te indikacije pristranskosti ni nihče komentiral, niti Dressel in Farid (2018) niti Rudin (2019) niti v drugih člankih, ki preučujejo COMPAS.

Iz predstavljenih rezultatov je mogoče sklepati, da je napovedovanje tveganja povratništva kljub obsežnim razpoložljivim informacijam o obtožencih izjemno zahtevno, saj boljše napovedne točnosti, kot kaže, ni mogoče doseči. Obenem se zdi, da je skoraj vse doseženo z dvema ali s tremi najbolj koristnimi atributi in da preostalih več kot 130 atributov ne prispeva k izboljšanju napovedi. V skladu s tem nekateri avtorji (Angwin et al. 2016; Dressel in Farid 2018) zaključujejo, da uporaba strojnega učenja v pravosodju nima obetavne prihodnosti. Tak zaključek pa je prenagel in preveč poenostavljen, na kar opozarja Spielkamp (2017). V številnih drugih aplikacijah, kot je medicinska diagnostika, je strojno učenje v večini primerov že preseglo napovedno točnost strokovnjakov, kar so potrdili številni eksperimenti, npr. Cestnik, Kononenko in Bratko (1987).

Različna mnenja o pristranskosti in uporabnosti sistema COMPAS kažejo na pomanjkanje splošno sprejetih operativnih definicij pristranskosti in pravičnosti v strojnem učenju. Problem lepo ponazarja pogosto citirani članek (Mehrabi et al. 2021), ki preučuje številne relevantne definicije pristranskosti, vendar ne ponudi skupne sinteze, ki bi zmanjšala konceptualno kompleksnost in omogočila praktičen pristop k problemu pristranskosti. Članek povzroči dodatno zmedo, ko hitro odpravi COMPAS kot očitno pristranski in neuporaben, pri čemer ne upošteva argumentov Floresa, Bechtela in Lowenkampa (2016).

Formalne definicije pravičnosti in izzivi za implementacijo v praksi

V splošnih medijih se strojnemu učenju pogosto očita pristranskost bolj po občutku, brez natančne opredelitve matematično preverljivih kriterijev, na podlagi katerih bi bilo pristranskost mogoče dokazati. Izjave, kot sta: **»**Sistem se je v sodstvu izkazal kot pristranski do temnopoltnih obtožencev**«** (Angwin et al. 2016) ali **»**Sistem je pri ocenjevanju kandidatov za zaposlitev pristranski do žensk**«** (Blanzeisky in Cunningham 2021) uporabljajo splošne izraze, kot so **»**pristranskost algoritmov**«**, **»**pristranskost strojnega učenja**«**, **»**pristranskost umetne inteligence**«**. V mnogih primerih so bile te izjave opremljene s preprostimi, a pogosto preveč poenostavljenimi razlagami, na primer: **»**Sisteme strojnega učenja razvijajo skoraj izključno belopolti moški, zato …**«**

Danes je jasno, da problem pristranskosti ni tako trivialen. Pretirano poenostavljene razlage so vse redkejše, prav tako pa postaja očitno, da izraz »pristranskost algoritmov« ni primeren, saj ustvarja napačen vtis, da algoritmi lahko imajo zle namene in da ne delujejo na podlagi matematičnih in statističnih principov (Hardt, Price in Srebro 2016). Glavni cilj metod strojnega učenja je vedno odkrivanje zakonitosti v realnem svetu na podlagi razpoložljivih podatkov. Težava nastane, če v realnem svetu že obstajajo pristranske prakse. Podatki, zbrani v takem okolju, odražajo pristranskost, ki jo algoritem za učenje zazna in reproducira. Če nato rezultate, dobljene iz pristranskih podatkov, znova uporabimo v realnem svetu, s tem reproduciramo in utrjujemo že obstoječo pristranskost (ibid.).

Kljub vsemu še vedno ni dovolj natančno definirano, kaj pristranskost pravzaprav je. Pogosto gre za vtis pristranskosti, pri čemer se predsodki za ali proti posamezniku ali skupini kažejo na način, ki je zaznan kot nepravičen (Ntoutsi et al. 2020).

Poglejmo, v čem so težave pri definiranju pristranskosti. Že na področju strojnega učenja se termin »pristranskost« uporablja v raznih pomenih. Izraz se v strojnem učenju nanaša na več fenomenov (Hellström, Dignum in Bensch 2020):

- *Induktivna pristranskost*: To je princip, po katerem se algoritem odloči za eno izmed številnih možnih hipotez, ki so vse na nek način utemeljene z učnimi podatki. Gre za nabor (eksplicitnih ali implicitnih) predpostavk algoritma, katerih cilj je posplošiti niz opazovanj (učnih podatkov) v splošni model domene (Hüllermeier, Fober in Mernberger 2013). Ta vrsta pristranskosti je nepogrešljiv mehanizem, ki omogoča strojno učenje, zato je v principu pozitivna komponenta, brez katere strojno učenje sploh ni mogoče. Primer take pristranskosti je načelo Occamove britve, ki pravi: če imamo na voljo dve razlagi zbranih podatkov, ki obe enako dobro razložita podatke, je bolj smiselno izbrati preprostejšo razlago (Gordon in Desjardins 1995; Hellström, Dignum in Bensch 2020; Ntoutsi et al. 2020). Čeprav ima izraz pristranskost pogosto negativen prizvok, je induktivna pristranskost pozitivna in celo nujna komponenta strojnega učenja, kar lepo razložijo Gordon in Desjardins (1995) in osnovni učbeniki UI.
- *Pristranskost v učnih podatkih*: Pristranskost, ki odraža obstoječe pristranskosti v ustaljenem odločanju na danem področju uporabe (npr. pristranskost strokovnjakov v dejanski sodni praksi, iz katere so pridobljeni učni podatki) (Blanzeisky in Cunningham 2021; Dastin 2018). Alelyani (2021), poudarja, da je taka pristranskost pogosto posledica kognitivne pristranskosti, ki je naravni fenomen človeškega mišljenja. Človeški možgani namreč filtrirajo ogromne količine informacij in ohranjajo le tiste, ki so za nas relevantne. Ker so algoritmi učeni na podatkih, ki odražajo človeško vedenje, pogosto reproducirajo kognitivno pristranskost, ki jo vsebujejo vhodni podatki. Avtorji to vrsto pristranskosti poimenujejo različno: Blanzeisky in Cunningham (2021) jo označujeta kot negativno dediščino (»negative legacy«), medtem ko jo Hellström, Dignum in Bensch (2020) imenujejo zgodovinska pristranskost (»historical bias«).
- Pristranskost, ki izhaja iz neprimernega postopka zbiranja podatkov oz. *vzorčenja* (Hellström, Dignum in Bensch 2020): Ta vrsta pristranskosti se pojavi, kadar je npr. za določeno skupino ljudi na voljo bistveno manj primerov kot za druge

skupine. V skladu z matematično utemeljenimi statističnimi in verjetnostnimi načeli lahko nekatere skupine, tipično manjšinske, izpadejo kot diskriminirane (včasih celo v pozitivnem smislu!) preprosto zato, ker metode za ocenjevanje verjetnosti upravičeno ocenijo te verjetnosti drugače, kadar je na voljo manj podatkov. Tako pristranskost Blanzeisky in Cunningham (2021) imenujeta podcenjevanje (»underestimation«). Sun, Nasraoui in Shafto (2020) opozarjajo na ključno vprašanje o izvoru učnih podatkov. Tradicionalno so se algoritmi pri učenju zanašali na zanesljive oznake, ki so jih določili strokovnjaki. Danes pa se mnogi algoritmi učijo iz podatkov, ki izvirajo iz širše družbe, pri čemer so oznake in vzorci pogosto pristranski.

Prej omenjeni viri pristranskosti so razmeroma splošno sprejeti. Kljub temu ostaja izziv natančno definirati kriterije, ki bi objektivno določili, ali je dani sistem pristranski, ter omogočili tudi kvantitativno vrednotenje te pristranskosti. Obstajajo razne matematične mere, ki se zdijo smiselne, vendar so med seboj neskladne. Posledično za zdaj še ne obstaja preprosta in splošno sprejeta mera pristranskosti.

To kompleksnost dobro ponazarja izčrpni pregled različnih definicij pravičnosti kot nasprotja pristranskosti (Ntoutsi et al. 2020). Pravičnost lahko opredelimo na več načinov, na primer: »pravičnost na podlagi zavedanja« (»*fairness through awareness*«), pri čemer algoritem velja za pravičnega, če za podobne posameznike poda podobne napovedi; enakost obravnave (»*treatment equality*«), pri čemer se zahteva enako razmerje med FNR in FPR v obeh skupinah glede na zaščiteni atribut (npr. raso); poštenost v relacijskih domenah (»*fairness in relational domains*«), pri čemer poleg atributov posameznikov upoštevamo še družbene, organizacijske in druge povezave med njimi.

Berk et al. (2021) ugotavljajo, da so različne formalne definicije pravičnosti lahko med seboj nekompatibilne, kar vpliva na zanesljivost ocen tveganja in predlagane algoritmične rešitve. Avtorji ugotavljajo, da, razen v trivialnih primerih, ni mogoče hkrati optimizirati točnosti in pravičnosti ter zadostiti vsem vrstam pravičnosti hkrati. Ta problem podrobneje preučijo Kleinberg, Mullainathan in Raghavan (2016). Definirajo tri osnovne in na videz očitne pogoje, ki jih mora sistem izpolnjevati, če naj bo nepristranski (pravičen). Presenetljivo se izkaže, da ti trije pogoji ne morejo biti izpolnjeni hkrati, razen v posebnih, trivialnih primerih, ki pa so za prakso nezanimivi. Torej so že te tri osnovne zahteve skupaj neuresničljive. Te tri zahteve so:

1. Kalibracija ocen verjetnosti. Če algoritem identificira skupino oseb, za katero ocenjujejo, da imajo določeno verjetnost za pripadnost pozitivnemu razredu, mora ta delež dejansko pripadati pozitivnemu razredu. Enak pogoj mora biti izpolnjen tudi za vse skupine oseb, ki se razlikujejo glede na zaščiteni atribut, kot je rasa ali spol. Ocene morajo odražati to, kar naj bi pomenile, in biti neodvisne od skupine glede na zaščiteni atribut, v katero posameznik spada.
2. Ravnotežje pozitivnega razreda. Povprečje »ocen tveganja« (stopnja pripadnosti pozitivnemu razredu) oseb pozitivnega razreda mora biti enako za vse skupine. V primeru sistema COMPAS bi to pomenilo, da bi morali imeti belopolti in temnopolti obsojenci, ki pripadajo pozitivnemu razredu, primerljive ocene tveganja.

3. Ravnotežje negativnega razreda. To načelo dopolnjuje prejšnje načelo pozitivnega razreda in pomeni, da morajo biti povprečne ocene stopnje pripadnosti negativnemu razredu za osebe v negativnem razredu enake za vse skupine.

Avtorji matematično dokažejo, da so te tri zahteve, čeprav si prizadevajo za isti cilj zmanjševanja pristranskosti, med seboj nekompatibilne, razen v posebnih primerih.

Kadar se pojavi pristranskost, se pojavi vprašanje, kako jo odpraviti. Obstaja več pristopov, izmed katerih sta najočitnejša (a) prepovedana uporaba »zaščitenih atributov« in (b) obratna diskriminacija. Tipična zaščitena atributa sta rasa in spol. Princip zaščitenih atributov pomeni, da algoritmu učenja prepovemo uporabo teh atributov pri odločanju o klasifikaciji posameznega primera. Ta pristop pogosto ni učinkovit, saj algoritem učenja uspe rekonstruirati vrednosti zaščitenih atributov na podlagi drugih, nezaščitenih atributov, ki korelirajo z zaščitenimi. Na podlagi podatkov o izobrazbi ali lokaciji prebivališča lahko na primer algoritem sklepa o rasi osebe (Hardt, Price in Srebro 2016).

Princip obratne diskriminacije temelji na tem, da deprivilegiranim skupinam namenoma damo določeno prednost, s čimer naj bi odpravili učinke diskriminacije. Čeprav je ta ukrep očitno dobronameren, v praksi uvaja novo obliko nepravičnosti, ki je za nekatere vprašljiva (Alelyani 2021). Taka nepravičnost (obratna diskriminacija) je sicer lahko upravičena, vendar ne z vidika pravičnosti, temveč z vidika »višjih« vrednot, kot so popravljanje zgodovinskih krivic in doseganje dolgoročne pravičnosti na podlagi začasne nepravičnosti. Gre torej za strateško uresničevanje družbeno sprejetih vrednot, ki pa v praksi zaradi zgodovinskih razlogov in vztrajnosti zahtevajo daljše časovno obdobje za implementacijo. Ostaja težavno vprašanje, do kolikšne mere je obratna diskriminacija smiselna, kar bi moralo biti določeno z demokratično sprejetim družbenim konsenzom, formaliziranim z ustreznimi zakonskimi rešitvami za vsak specifičen primer.

V praksi se reševanja pristranskosti lotimo znotraj treh faz strojnega učenja: 1) predprocesna faza, v kateri povečamo vzorec manjšine za bolj uravnoteženo predstavitev podatkov; 2) medprocesna faza, v kateri dodajamo omejitve, s katerimi kompenziramo za neenakomeren vzorec in 3) poprocesna faza, v kateri prilagajamo mejne vrednosti odločitve, predvsem za manjšinske skupine (Blanzeisky in Cunningham 2021; Mehrabi et al. 2021; Ntoutsi et al. 2020).

Pri razvoju metod in orodij je ključno, da se zavedamo potencialnih pasti (Alelyani 2021; Holsinger et al. 2018). Avtorji opozarjajo, da lahko nekatere rešitve nenamerno vodijo v nove oblike nepravičnosti. Hüllermeier, Fober in Mernberger (2013) izpostavijo pogost stranski učinek, ko poseganje v učne podatke privede do izgube pomembnih povezav med atributi ali celo do poslabšanja delovanja modela, merjenega s kazalniki, kot sta točnost in ocena F1. Kot primarni vzrok pristranskosti avtorji identificirajo predhodne odločitve, ki so generirale same učne podatke. Predlagajo taktiko Fair-SMOTE, ki omogoča zmanjšanje pristranskosti, obenem pa ohranja (ali celo izboljšuje) delovanje sistema. Ključni korak je mutacija podatkov, z enakomerno ekstrapolacijo za vse spremenljivke. Fair-SMOTE identificira neuravnoteženost podatkov in napačno označene podatkovne točke. Rezultat je generiranje bolj poštenih rezultatov. Avtorji zavračajo pomisleke (Berk et al. 2021), da je cena poštenosti

poslabšanje delovanja sistema. Zaključijo, da je namesto slepe uporabe optimizacijskih metod pomembneje upoštevati specifičnosti domene in to znanje uporabiti za izboljšanje v pogledu pravičnosti.

Zaključek

Pristranskost v nekaterih ključnih aplikacijah strojnega učenja je postala popularna in kontroverzna tema. V razpravah je nejasnost, ki izvira iz dejstva, da večina ljudi razume pojem pravičnosti in pristranskosti intuitivno. Pri tem pravičnost doživljamo na različne načine in v podrobnostih ni popolnega soglasja. Tako tudi ni soglasja o tem, kakšen naj bi bil jasen, matematično formuliran kriterij, s katerim bi brez dvoma kvantificirali pristranskost konkretnega sistema. Obstaja veliko kontroverznih vprašanj in odprtih tem, v zvezi s katerimi ni strinjanja. Ni soglasja o izvoru pristranskosti, niti o tem, katera orodja oz. metode so za soočanje s pristranskostjo najprimernejše.

Poštenost v strojnem učenju naj bi pomenila produkcijo odločitev, s katerimi bi bila družba zadovoljna. Problem nastane, ker si v tem nismo enotni. Primer COMPAS ponazarja, kako nujna je zedinjenost. COMPAS so analizirali različni strokovnjaki, katerih ugotovitve so si nasprotujoče. Nekateri pravijo, da je COMPAS pristranski, medtem ko drugi menijo, da ni. Spielkamp (2017) verjame, da imajo prav vsi, saj vsakdo razume pravičnost na svoj način. Študija Kleinberga, Mullainathana in Raghavana (2016) je še posebej zanimiva, saj avtorji matematično dokažejo, da so določene definicije pravičnosti med seboj nezdružljive, čeprav se na prvi pogled vse zdijo enako pomembne in nujne.

Za kakovostno obravnavanje pristranskosti so pomembni jasna opredelitev zaželenih družbenih vrednot, upoštevanje zgodovinskega okvirja in resničnega stanja sveta ter tudi izobrazba o vsaj osnovnih principih strojnega učenja. Šele na temelju tega znanja lahko kakovostno artikuliramo svoja pričakovanja. Potem se lahko odločimo, kateri koncept pravičnosti ustreza v specifičnih kontekstih. Corbett-Davies et al. (2018) menijo, da ima lahko v določenih kontekstih delovanje v skladu s popularnimi koncepcijami pravičnosti brez globljega razumevanja domene celo nasprotne učinke. Avtorji v kontekstu razvoja algoritmov predlagajo, da se namesto osredinjanja na aksiomatska razumevanja pravičnosti raje osredinimo na njihove posledice, ki so močno odvisne od konteksta in domene.

V literaturi na splošno ni dobro predvideno, kako velik izziv bo v praksi reševanje problema pristranskosti. Pričakovanja glede vrednot bo treba namreč natančno opredeliti z ustreznimi zakonodajnimi rešitvami, ali naj bo na primer zaradi zgodovinskih krivic v konkretni aplikaciji realizirana obratna pristranskost in v kolikšni meri. Take formulacije bodo morale biti bolj tehnične, kot je običajno v zakonih in drugih predpisih, saj bodo osnova za konkretno implementacijo v algoritmih UI. Jasno je, da problem pristranskosti ni (zgolj) tehnološke narave in da morajo zato učinkoviti pristopi k reševanju problema vključevati tudi širši nabor ukrepov (Ntoutsi et al. 2020). Prizadevati si moramo za interdisciplinarno raziskovanje, pri katerem bodo inženirji na področju UI tesno sodelovali s strokovnjaki s področij etike in odločanja (Yu et al. 2018).

Za ustrezno razumevanje in ukrepanje na tem področju je ključna kakovostna splošna izobrazba. Pomanjkanje se namreč kaže v načinu poročanja, odzivanju jav-

nosti in zmedenosti strokovnjakov. Različni algoritmi postajajo neizogiben del vsakdanjosti. Nesprejemljivo je, da o njih ne samo, da vemo premalo, ampak imamo celo napačne predstave. Po drugi strani imamo tudi institucije, ki delovanja sistemov UI prav tako ne razumejo, hkrati take sisteme naročajo in jih nato uvajajo v svoje procese odločanja. Institucije pogosto nimajo dovolj znanja in virov, da bi znale specificirati primerna algoritmična orodja. Nujno je izobraziti ljudi, da bodo sposobni ustrezno artikulirati, kaj naj bi algoritem res meril in kateri kriteriji morajo biti izpolnjeni, da bo algoritem pravičen (Courtland 2018).

Razložljivost v umetni inteligenci

Problem razložljivosti v umetni inteligenci

Uspehi in uporabnost tehnologije strojnega učenja so privedli do splošne uporabe te tehnologije. Sistem se iz danih primerov nauči dajati napovedi in odločitve, ki so tipično kakovostne glede na točnost, vendar ne tudi vedno povsem zanesljive. V nekaterih primerih uporabe, kot je npr. zdravstvo, imajo lahko napačne odločitve zelo pomembne ali celo usodne posledice.

Na takih področjih uporabe ni pomembna le točna napoved sistema in je nujno, da uporabnik lahko razume in preveri odločitve UI (Barredo Arrieta et al. 2019). Za to je pomembno, da ima model strojnega učenja zmožnost razlage, kako je prišel do odločitve. Nekatere metode strojnega učenja brez težav tako razlago tudi res omogočajo. Nekatere druge metode, npr. posebej zmogljive nevronske mreže in globoko učenje (LeCun, Bengio in Hinton 2015), pa slovijo po nerazumljivosti. Napovedi, ki jo take metode dajejo, so tipično rezultat interakcij med ogromnim številom numeričnih parametrov (reda bilijon), kar daleč presega človekove sposobnosti razumevanja. Znan je t. i. problem *črnih škatel* (»*black box problem*«), ki pomeni, da delovanje modelov strojnega učenja ostaja za uporabnike nerazumljivo. Prav pomanjkanje razumevanja omejuje nadaljnjo in bolj praktično uporabo modelov v mnogih pomembnih domenah odločanja.

Potreba po razlagi je vodila v razvoj tehnik in pristopov t. i. razložljive umetne inteligence (XAI – »*eXplainable Artificial Intelligence*«), ki se posveča nalogi razlaganja kompleksnih modelov strojnega učenja. Nove učinkovite metode razlage so posebej pomembne na področju nevronskih mrež in globokega učenja. Vendar je treba ugotoviti, da je tehnologija globokega učenja prešla v splošno uporabo, še preden je bil razvit kakovosten konceptualni okvir, ki bi omogočal jasno razumevanje delovanja in odločitev globokih nevronskih mrež.

V nadaljevanju tega poglavja sledita pregled trenutnega stanja metod XAI in analiza pomanjkljivosti.

Kaj sploh je razlaga?

Razložljivost je izmuzljiv pojem ne samo na področju umetne inteligence (UI), pač pa širše na področju filozofije in drugih družboslovnih znanosti. Na področju UI se operira s koncepti, kot so vzročnost, informativnost, razumevanje, gotovost, zaupanje, transparentnost ipd. (Barredo Arrieta et al. 2019). Termin »razložljiva umet-

na inteligenca« je leta 2019 kot del svojega programa uporabila DARPA (Gunning et al. 2021). Od takrat je postal zelo popularen, ne gre pa za nov pojav. Kvečjemu gre za novo imenovanje dolgoletnih prizadevanj, med katerimi se raziskovalci trudijo prebiti do odgovora na vprašanje, zakaj je sistem prišel do določene napovedi (Holzinger et al. 2022). Prvi je verjetno opozoril na problem razložljivosti Donald Michie, in to že precej prej, kot je to formuliral v svoji objavi (Michie 1988).

Najsplošneje bi lahko razložljivost v UI opredelili kot razlago, zaradi katere je delovanje modela razumljivejše. Seveda je to zelo splošna opredelitev, v poskusih natančnejšega definiranja pa si raziskovalci niso enotni. Barredo Arrieta et al. (2019) navajajo, da mora model nuditi razlago za svoje delovanje in napovedi v obliki vizualizacije pravil in vpogleda v spremenljivke, ki bi lahko povzročile perturbacije modela. Ribeiro, Singh in Guestrin (2016) besedo "razložiti"opredeljujejo kot predstaviti besedilne ali vizualne elemente, ki omogočajo kvalitativno razumevanje odnosa med komponentami in napovedjo modela.

Ena od nekonsistentnosti v literaturi iz XAI je uporaba pojma interpretabilnost, ki je včasih sinonim razložljivosti, drugič ločen pojem, tretjič ena od kategorij razložljivosti. Barredo Arrieta et al. (2019) interpretabilnost razumejo kot pasivno, razložljivost pa kot aktivno lastnost modela. Interpretabilni so modeli, ki so razumljivi že sami po sebi (npr. odločitvena drevesa), razložljivi pa tisti, za katere obstajajo metode, ki generirajo ustrezno razlago.

Očitno je pomanjkanje konsenza o glavnih konceptih. Problem je, da vsaka definicija nastopa znotraj specifičnega konteksta, odvisnega od naloge, sposobnosti in pričakovanj raziskovalca. Opredelitve razložljivosti so tako pogosto vezane na specifično domeno. Posledično področje XAI še ni enotno glede definicije razlage, specifičnih ciljev in kriterijev, ki naj bi jim zadostovali modeli, da bi bili razumljivi (ibid.).

Metode razlag

Obstajajo številne metode razlag. Problem pa nastane pri njihovi klasifikaciji, ker ima praktično vsak avtor specifično definicijo razložljivosti, iz katere izhaja.

Ena splošnih kategorizacij je delitev na lokalne in globalne razlage. Lokalne so razlage, središčene okoli posameznega primera, pri čemer pa ostane delovanje modela kot celote nepojasnjeno. Na drugi strani globalne razlage omogočajo razumeti celoten model, so pa pogosto osnovane na približnih vrednostih (Alvarez-Melis in Jaakkola 2018; Hall, Ambati in Phan 2017; Ignatiev, Narodytska in Marques-Silva 2019; Yeh in Ravikumar 2021).

Druga splošna delitev je na modelno odvisne in modelno neodvisne (»agnostične razlage«, »*model-agnostic*«) razlage. Slednje s tehnikami, kot so relevantnost atributov, vizualizacija in poenostavitev, pridobijo določene informacije o postopku napovedovanja in so uporabne za vsako vrsto modela (Barredo Arrieta et al. 2019). Modelno odvisne razlage so uporabne zgolj za specifične vrste modelov (npr. maksimizacija aktivacije nevronov, ki jo opišejo Guidotti et al. (2018)) (Hall, Ambati in Phan 2017).

Barredo Arrieta et al. (2019) ločijo besedilne, vizualne, lokalne razlage, razlage s primeri, poenostavitvijo in z relevantnostjo atributov. Chen et al. (2022) opredelijo tri glavne kategorije razlag: osnovane na funkciji, na primerih in pojasnjevanju atri-

butov. Yeh in Ravikumar (2021) ločita razlage atributov in razlage primerov. Ali et al. (2023) razlage razdelijo glede na uporabljeno metodologijo in ločijo med razlagami, ki slonijo na vzvratnem razširjanju (»*backpropagation*«), in razlagami s perturbacijami. Guidotti et al. (2018) ločijo: 1) odločitvena drevesa; 2) razlage, osnovane na pravilih; 3) razlage pomembnosti atributov, ki predstavijo težo in pomembnost atributov, ki jih je pri svoji napovedi upošteval model. Primer je znana metoda LIME (»*Local Interpretable Model-agnostic Explanation*«), primerna predvsem za razlago klasifikacije besedil in slik (Ribeiro, Singh in Guestrin 2016); 4) zemljevidi pomembnosti, ki izpostavijo ključne vidike predmeta, ki je analiziran. Primer je metoda CAM (npr., da gre za sliko z umivanjem zob, razložimo tako, da posebej poudarimo dele slik, na katerih se zobna ščetka približa ustom) (Zhou et al. 2016); 5) PDP (»*Partial Dependence Plot*«), pri čemer grafično prikažemo odnos med odločitvijo modela in vhodnimi podatki; 6) razlaga s prototipi, pri kateri z napovedjo dobimo primer, podoben našemu; 7) maksimizacija aktivacije, pri kateri opazujemo, kakšni vzorci vhodnih podatkov maksimizirajo aktivacijo določenega nevrona oz. nivoja.

Ignatiev, Narodytska in Marques-Silva (2019) predstavijo pojem formalne razložljivosti, zasnovan na logiki, pri čemer so razlage posledično zanesljivejše in globalno veljavne. Pristop temelji na računanju t. i. glavnih implikantov (»*prime implicant*«), kar omogoča kompaktno logično predstavitev delovanja modela.

Razni kriteriji dobre razlage

Če je eden od ključnih ciljev XAI izboljšanje zaupanja v sisteme UI, je nujno, da se pozornost usmeri k uporabnikom teh sistemov (Zhang et al. 2021). Dobre razlage bodo tiste, ki bodo upoštevale, komu so namenjene (Barredo Arrieta et al. 2019). To pomeni upoštevanje predznanja, ki ga imajo uporabniki. Opazen je trend, da razvijalci metod razlag tega ne upoštevajo dovolj. Ribera in Lapedriza (2019) opredelita tri skupine uporabnikov (razvijalci in raziskovalci, eksperti in laiki), ki zahtevajo različne vrste razlag.

Raziskovalci predlagajo različne kriterije za dobro razlago. V nadaljevanju navajamo nekaj primerov kriterijev.

Alvarez-Melis in Jaakkola (2018) opredelita tri kriterije:

- eksplicitnost: razlaga je takojšnja in razumljiva;
- t. i. zvestoba (»*faithfulness*«): ocene relevantnosti odražajo resnično pomembnost;
- stabilnost: za podobne vhodne podatke veljajo podobne razlage.

Chen et al. (2022) poudarjajo:

- robustnost oz. občutljivost: ustrezna sprememba razlage v primeru spremembe vhodnih podatkov;
- zvestoba: razlaga ponazarja dejansko odločanje modela;
- kompleksnost: kognitivni napor za razumevanje razlage;
- homogenost: zmožnost za pravilno razlago delovanja modela glede na različne skupine (v praksi se to po navadi nanaša na skupine, ki se razlikujejo glede na občutljive, zaščitene atribute).

Amgoud in Ben-Naim (2022) opredelita štiri aksiome, ki naj bi jim zadostile dobre razlage:

1. morajo biti informativne;
2. ne smejo vsebovati nepotrebnih informacij;
3. razlage razredov morajo pojasniti posamezne primere, hkrati pa morajo biti splošno uporabne;
4. razlaga mora vsebovati samo informacije, ki vplivajo na napoved.

Ocenjevanje razlag

Ocenjevanje razlag je najmlajše področje s široko paleto pristopov (Ribera in Lapedriza 2019). Za razliko od točnosti je kriterije, kot so varnost, pristranskost in razložljivost, težje kvantificirati (Doshi-Velez in Kim 2017).

Ocenjevanja se (najsplošneje) lahko lotimo na dva načina: 1) s človeškim ocenjevanjem ali 2) z uporabo računskih metod, ki merijo, kako dobro razlaga dejansko razloži delovanje modela. Glavna razlika med pristopoma je, da so računske metode objektivnejše, vendar pa upoštevajo človeški faktor le posredno. Drugače rečeno, ne kvantificirajo nujno človeškega razumevanja. Prednosti človeške ocene sta subjektivnost in večja deskriptivnost. Očitni pomanjkljivosti sta manjša ponovljivost in večja odvisnost od specifične naloge (Mohseni, Block in Ragan 2021).

Hsieh et al. (2021) in Zhang et al. (2021) predstavijo matematično ocenjevanje razlag na podlagi analize robustnosti. Matematično opredeljena mera nezvestobe ponazarja, kako dobro se razlaga ujema z modelom (Yeh in Ravikumar 2021).

Edmonds et al. (2019) so izvedli eksperiment, s katerim so preverili, kakšne razlage so pri ljudeh vzbudile največje zaupanje v robota, ki je odprl stekleničko. Robot se je naučil odpirati stekleničke iz človeških demonstracij, pri čemer je bilo ključno učenje zaporedij položaja rok in potrebne sile. Z rokavico s senzorji so zajeli podatke o sili in položaju rok v 64 človeških demonstracijah s tremi različnimi steklenič kami. Sledilo je kompleksnejše učenje, da bi bil robot svoje znanje zmožen posplošiti. Implementiran je bil haptični model, ki je robotu pomagal določiti potrebno silo, čeprav nima človeških rok. Ker odpiranje stekleničke poteka v več korakih (potiskanje, odvijanje itd.), je bil uporabljen tudi algoritem simboličnega planiranja, ki upošteva pravila o zaporedju potrebnih akcij, ki omogočajo druga drugo. S kombinacijo takega učenja je robot postal precej dober v odpiranju novih stekleničk. Udeleženci so bili razdeljeni v pet skupin. Vsaka je videla posnetek robota, ki opravlja nalogo, ter eno od možnih razlag: 1) simbolično: v realnem času so udeleženci videli z eno besedo opisano akcijo, ki naj bi razlagala, kaj robot na posnetku dela (npr. »*approach – grasp – push – twist – ungrasp – move – grasp – push …*«); 2) besedilno: po ogledu posnetka robota so udeleženci prebrali kratko besedilo o tem, kako je robot opravil nalogo (npr. »*I succeeded to open the bottle because I pushed on the cap three times and twisted the cap twice«*); oz. 3) haptično razlago: vizualizacija sile prijema v vsakem trenutku odpiranja stekleničke oz. kombinacijo haptične in simbolične razlage. Največ zaupanja je spodbudila simbolična razlaga. Ta rezultat nakazuje, da so principi simboličnega planiranja koristni za razlago robotskega delovanja, tudi če niso nujni za planiranje rešitve naloge.

Mohseni, Block in Ragan (2021) so izvedli eksperiment, med katerim so udeleženci označili relevantna področja slike, ki so po njihovem mnenju bila najbolj reprezentativna za določen razred objektov (npr. mačka in pes). Rezultat je zemljevid pomembnosti, ki prikazuje območja slike, ki so jim udeleženci posvečali največ pozornosti. Te rezultate so primerjali z zemljevidi pomembnosti metode Grad-CAM. Zemljevidi so si podobni, vseeno pa je statistično testiranje pokazalo pomembne razlike. Distribucija relevantnih atributov je bila pri metodi Grad-CAM bolj uniformna. Udeleženci so v primeru živih bitij kot ključne pogosteje označevali obraze. Prav to so ugotovitve, ki so lahko v oporo pri razumevanju, kako dobre so razlage.

Nekatera odprta vprašanja

V tem razdelku opozorimo na nekatere razmeroma manj raziskane probleme in tudi na manj uporabljene pristope za XAI.

Tehtanje med točnostjo in razložljivostjo

Rudin (2019) v članku z zgovornim naslovom »*Stop explaining black box ML models for high stakes decisions and use interpretable models instead*« izraža determinirano stališče. Zavzema se za uporabo metod učenja, ki dajejo naučene modele, ki so sami po sebi razumljivi. Za take veljajo npr. odločitvena drevesa. Nasprotuje metodam učenja, katerih rezultati so v principu težko razumljivi. Med te štejemo posebno metode globokega učenja, ki sicer dosegajo visoko napovedno točnost v primerjavi z drugimi metodami učenja, toda ne zastonj: vsaj za ceno razumljivosti in potrebnega velikega števila podatkov za učenje. Pri tem gre Rudin morda res predaleč z optimističnim stališčem, ki implicitno predpostavlja možnost izgradnje elegantnih in razumljivih modelov za vsako problemsko domeno, s čimer zadene ob princip kompleksnosti Kolmogorova (Kolmogorov 1963).

Glede možnosti obstoja enostavnih modelov in razlag velja vsaj ena teoretična omejitev, ki jo definira kompleksnost Kolmogorova, ki določa, koliko spominskega prostora potrebujemo za najkrajši možni zapis danega objekta v računalniku. Obstajajo zapleteni objekti (torej tudi zapleteni napovedni modeli), ki jih niti teoretično ni mogoče predstaviti na kratek način. V takih primerih tudi razlaga ne more biti kratka in preprosta. Res pa je, da smo v praksi še zelo daleč od te teoretično dosegljive meje, torej imamo veliko prostora za izboljšanje. Ko zadenemo ob zid Kolmogorova, pa je še vedno možen kompromis, da za boljšo razložljivost žrtvujemo nekaj točnosti (Bratko 1997). Primer tehnične izvedbe tega tehtanja med točnostjo in razumljivostjo v učenju odločitvenih dreves sta razvila Bohanec in Bratko (1994).

Navezava razlage na uporabnikovo predznanje

Ali bo razlaga dobra, je odvisno od uporabnika razlage, konkretno od uporabnikovega predznanja o problemski domeni. Če je to kakovostno, zadošča en namig. Če je razumevanje domene slabo, mora biti razlaga podrobna in daljša. Tudi formulacija razlage je odvisna od obstoječega znanja na obravnavanem področju. Celo povsem pravilna in jedrnata razlaga je za eksperta na področju uporabe lahko nesprejemljiva in nenaravna. Kot primer omenimo, da so se nekateri primeri razlag, ki jih generirajo naučeni modeli v medicinskih domenah, kljub diagnostični točnosti zdravniku zdeli

povsem nenaravni (Bratko 1997). V enem od primerov je sistem razložil, da gre za vnetni revmatizem, ker ima pacient med drugim več kot dva prizadeta sklepa na roki. To diagnostično pravilo je dejansko točno. Vendar pa je zdravnik vztrajal, da mora imeti pacient prizadete sklepe na vseh petih prstih na roki, ker vnetni revmatizem tipično vpliva na vse sklepe. Ekspertno mnenje je bilo v tem primeru zelo jasno, čeprav je res, da bo pravilo vodilo do pravilne diagnoze v vsakem primeru, če je število vnetih sklepov karkoli med 2 in 5. Ustreznost razlage je odvisna ne le od klasifikacijske točnosti, temveč (tudi) od predznanja, ki ga ima uporabnik o tej obliki revmatizma.

Obstoječe metode razlage ta vidik povečini ignorirajo. Problem je tudi v tem, da ne omogočajo naravne uporabe predznanja. V tem pogledu je zelo obetaven pristop k strojnemu učenju t. i. induktivno logično programiranje (ILP), ki temelji na uporabi matematične logike. Že osnovna formulacija problema učenja v ILP vsebuje uporabo predznanja: dani so učni primeri E in predznanje BK (»*background knowledge*«), naloga učenja pa je sestaviti logično formulo H (hipoteza) tako, da primeri E logično sledijo iz BK in H.

Pristop ILP je skromno zastopan v obstoječih raziskavah iz strojnega učenja in razložljivosti. Lep primer njegove ustreznosti so raziskave, ki jih opisujejo Ai et al. (2023) in Muggleton et al. (2018). Te zasledujejo ne le osnovni cilj XAI (razlage odločitev strojnega učenja), temveč tudi cilj t. i. »ultrarazložljivosti«. Ta strožji kriterij strojnega učenja je definiral Michie (1988) (»*ultra strong criterion for ML*«). Strojno učenje je ultrarazložljivo, če je ne le razložljivo, temveč uporabniku omogoča tudi *operativno uporabo za lastno reševanje* novih problemov, npr. da strojno naučeno znanje lahko uporabi za lastno reševanje določenih matematičnih problemov ali igranje šaha.

Razlaga zaporedij odločitev v planiranju

Večina metod XAI generira razlago *posameznih* odločitev oz. klasifikacij. Pri razložljivem planiranju pa gre za razlago množice odločitev (npr. zaporedja akcij, ki robota vodi do cilja). Posebej za razlago planov se je formiralo področje razložljivega planiranja (Chakraborti, Sreedharan in Kambhampati 2020).

Razlaga planov je običajno zahtevnejša od razlage v klasifikacijskih problemih. Treba je razložiti, kako so posamezne akcije odvisne od drugih, da skupaj rešijo nalogo. Primer razlage zaporedja odločitev je razlaga šahovskih partij, pri čemer je treba razložiti celo zaporedje potez ali drevo odločitev, ki definira uspešno strategijo. Primer opisuje Bratko (2018) in vsebuje težko razumljive in briljantne poteze šahovskega programa AlphaZero.

Razlaga planov je aktualna tudi na področju vodenja sistemov. Lep primer razlage naučenega plana vodenja podajata Šoberl in Bratko (2023, 2025). Gre za klasično nalogo iz teorije vodenja sistemov: vodenje sistema voziček-palica. Na vozičku je vrtljivo vpeta palica. Palica je postavljena približno vertikalno, vendar se, če ne ukrepamo, prevrne na tla. S potiskanjem vozička levo oz. desno je treba loviti ravnotežni položaj palice okrog vertikale, obenem pa doseči, da se voziček horizontalno premakne iz začetnega položaja do danega cilja. Naučena strategija vodenja je lepo razložljiva. Najprej nekoliko presenetljivo potisnemo voziček v nasprotno stran od cilja, s čimer dosežemo, da se palica nagne proti cilju. To omogoči pospeševanje vozička

v smer proti cilju, hkrati pa se ohranja ravnotežje palice, ko je ta nagnjena naprej v smeri cilja.

Zaključek

Področje XAI se je v zadnjih petih do desetih letih močno razraslo. Mnogi zato predpostavljajo, da je bil to tudi začetek področja. V resnici je zavedanje pomembnosti, da naj bi bilo strojno učenje razložljivo, obstajalo že pred 40 leti (npr. Michie 1988). Že takrat so obstajale raziskave o razložljivih modelih. Kljub sedanji količini raziskav in nedvoumnih uspehih se še vedno kaže, da pogrešamo nekatere ključne odgovore. Npr., že pred desetletji se je v sklopu istih prizadevanj pojavilo zavedanje, da potrebujemo formalne mere za ocenjevanje kakovosti razlag. Take sprejete mere še ni. Raziskovalci pri ocenjevanju razlag uberejo različne pristope, odvisne od raznih kriterijev (konteksta, domene, uporabnikov itd.).

Glede vprašanja, kaj je sprejemljiva razlaga, se v pomanjkanju splošnejših in principialnih kriterijev v sedanji praksi uporablja predpostavka, da so nekateri modeli razložljivi kar po definiciji, torej razložljivi sami po sebi. Mednje npr. navadno štejemo odločitvena drevesa ali pravila *če-potem*. Toda tudi ta kriterij je arbitraren. Kaj, če je odločitveno drevo zelo veliko, npr. da ima milijon vozlišč?

V tem razdelku smo opozorili tudi na počasen napredek pri razvoju metod za razlago zaporedij odločitev. Sem sodi razlaga planov za reševanje nalog, ki imajo eksplicitno definirane cilje. Plan je lahko zaporedje akcij ali pa tudi množica akcij, ki so delno urejene v času. Tu je treba razložiti tudi to, kako se akcije med seboj dopolnjujejo in na kakšen način skupaj dosežejo cilj. S tem so povezani izzivi sklepanja v velikih jezikovnih modelih, ki jih predstavimo v nadaljevanju.

Eden izmed možnih pristopov, ki upošteva principe planiranja v UI, je upoštevanje odvisnosti med akcijami. Nekatere akcije v planu neposredno dosežejo kakega od ciljev plana. Druge akcije pa ne dosežejo nobenega danega cilja neposredno, njihova funkcija je, da dosežejo pogoje, ki morajo biti uresničeni, da je možno izvesti druge akcije v planu. Taka razlaga plana je seveda povsem logična. Navadno pa vsebuje preveč podrobnosti. Če plan vsebuje nekoliko večje število akcij, npr. nekaj deset, postane tako podrobna razlaga spet težko razumljiva in za uporabnika neprivlačna. V tem primeru bi bilo treba za sprejemljivo razlago plan razbiti v hierarhično strukturo, definirano s podcilji plana. Odkrivanje smiselnih podciljev pa je lahko zelo težavno. Poseben izziv je, kako poiskati take podcilje, ki rezultirajo v za človeka čim naravnejši razlagi.

Sposobnost sklepanja v velikih jezikovnih modelih

V tem poglavju smo navedli nekatere etične vidike UI, ki naj bi zagotavljali zaupanja vredno UI. Podrobneje smo opisali stanje v zvezi z zagotavljanjem dveh od teh vidikov: pravičnost strojnega učenja in razložljivost v UI. Kljub napredku na teh dveh področjih ostajajo odprta oz. ne dovolj raziskana vprašanja. Poleg tega je za varno in zaupanja vredno UI pomembno tudi, da sistem UI deluje tehnično pravilno, npr. v tem smislu, da zanesljivo daje pravilne rezultate. V tem razdelku opozorimo na dve vprašanji v zvezi z generativno UI, ki zahtevata dodaten tehnični razvoj. To sta zmož-

nost sklepanja v velikih jezikovnih modelih ter problem, včasih poimenovan kot neke vrste »kanibalizem« (»*self consuming models*«).

En tehnični vidik, ki zadeva generativno UI kot trenutno najpopularnejši način uporabe UI, je povezan z zmožnostjo sklepanja v velikih jezikovnih modelih. Ta vidik generativne UI vzbuja dvome. Sposobnost sklepanja v popularnih jezikovnih modelih vzbuja vtis krhkosti in zelo težko je predvideti napake v sklepanju. To dejstvo je videti paradoksalno v luči dejstva, da sicer na področju UI že dolgo obstajajo zmogljive in zanesljive metode za avtomatsko sklepanje (npr. Russell in Norvig 2020; Poole in Mackworth 2023). Vendar kaže, da uporaba že obstoječih in zanesljivih metod ni preprosta v okviru generativne UI. Na probleme opozarjajo razni avtorji, npr. Bubeck et al. (2023) in Mirzadeh et al. (2024). Prizadevanja v smeri tehničnih izboljšav (npr. Plaat et al. 2024) v pogledu sposobnosti sklepanja so sicer privedla do izboljšav, vendar so te v pogledu sistematične zanesljivosti še vedno negotove. Dober primer je poskus z velikimi jezikovnimi modeli, ki ga je predlagal P. van Eecke (osebna komunikacija). Modelu zastavimo dobro znano miselno uganko »volk-koza-zelje«, ki je definirana tako: Imam volka, kozo in zelje. Če pustim volka brez nadzora s kozo, jo bo volk požrl. Podobno je s kozo in z zeljem. Kako naj prepeljem vse tri stvari varno čez reko z majhnim čolnom, tako da lahko vzamem v čoln s seboj le eno stvar? Tipičen jezikovni model nalogo takoj prepozna in tudi ponudi pravilno rešitev. Toda kaj se zgodi, če nalogo nekoliko spremenimo, npr. da je koza sita in jo zato lahko pustimo samo z zeljem? Ali je možna krajša rešitev? Tu modeli pogosto odpovedo, naredijo napačen logičen sklep in predlagajo nepravilno rešitev. Mirzadeh et al. (2024) predstavijo in analizirajo vrsto primerov v reševanju matematično logičnih nalog in značilnih spodrsljajev, ki so jih naredili jezikovni modeli. Res je, da so po tej objavi mnoge od teh tipov težav v jezikovnih modelih tudi že odpravili. Vendar je vtis, da gre bolj za lokalne popravke kot pa za sistematične izboljšave sklepanja.

Opozorimo na še en tehnični vidik, ki lahko postane kritičen v nadaljnjem razvoju generativne UI. Gre za pojav, ki ga primerjajo z neke vrste »kanibalizmom« med modeli generativne UI (»*self-consuming language models*«). Pri tem je mišljen verjeten razvojni scenarij tehnologije generativne UI, ko se bodo poznejše generacije modelov učile iz sintetičnih vsebin, in ne naravnih, človeških besedil in drugih vsebin. Po tem scenariju bodo prejšnje generacije modelov generirale sintetične vsebine, ki bodo na spletu na voljo za učenje bodočih generacij modelov. Pri tem lahko pride do naključnih napak, ki postopno privedejo do poslabševanja jezikovnih modelov poznejših generacij (Alemohammad et al. 2023; Briesch, Sobania in Rothlauf 2024).

Zaključna razprava

V tem poglavju smo podali pregled etičnih vidikov umetne inteligence ter podrobneje opisali predvsem tehnične vidike dveh od njih, pristranskosti in razložljivosti. Nismo pa se še dotaknili bolj globalnih vprašanj o tem, kako bo hitri razvoj UI vplival na družbo in človeštvo sploh.

Očitno je, da se zelo zmanjšuje pomen mnogih poklicev, pa tudi način človeškega dela v mnogih poklicih. Primer je prevajanje. Nekateri strokovnjaki UI so že pred

letom 1970 obljubljali, da bo kmalu na voljo kakovostno strojno prevajanje. Vendar se to ni uresničilo še nekaj desetletij in tako je strojno prevajanje dolgo veljalo za morda največjo neizpolnjeno obljubo UI. Ta situacija se je dramatično spremenila v vsega nekaj zadnjih letih. Za večino praktičnih potreb je prevajanje z UI postalo dovolj kakovostno in dosegljivo za drobec cene človeškega prevajanja. Profesionalni človeški prevajalci so res pomembni le še v zahtevnih primerih, ko gre pri prevajanju za specifične strokovne vsebine ali pa prevod zahteva literarno ustvarjalnost.

UI močno spreminja tudi izobraževalne procese in cilje izobraževanja. Ni več jasno, ali je treba, da pri učencih izobraževanje zagotovi sposobnost samostojnega pisnega izražanja, saj komunikacija poteka namesto na relaciji človek – človek vse bolj posredno na relacijah človek – računalnik – računalnik – človek. Zelo se spreminja tudi način raziskovalnega dela, od pregleda literature o sorodnih raziskavah do samega pisanja člankov. Ni več jasno, kaj je pri sodelovanju človek-UI prispevek človeškega avtorja in kaj računalnika.

Vendar v razpravah o teh spremembah vse bolj skrbi tudi najusodnejše vprašanje: Ali bo UI »prevzela nadzor«? Za to vprašanje se v angleščini navadno uporablja fraza »*Will AI take over?*« Lahko si je predstavljati, da se lahko z UI posamezniki ali skupine na neupravičen način dokopljejo do oblasti ali si ustvarijo druge neupravičene prednosti ali koristi. Vendar so take grožnje še vedno razmeroma obvladljive, saj so v očitnem nasprotju z obstoječimi zakoni in zato tudi pregonljive in jih je posledično vsaj v principu možno preprečiti z običajnimi sredstvi.

Obstajajo pa verjetnejši scenariji, da UI prevzame nadzor. Taki scenariji skrbijo vse več strokovnjakov UI in tudi drugih področij. Geoffrey Hinton, Nobelov nagrajenec leta 2024 za dosežke v učenju globokih nevronskih mrež, na primer ocenjuje, da je to skoraj neizogibno. Meni, da bi to lahko preprečili le z usklajenim ukrepanjem večine držav sveta, kar bi se lahko zgodilo, če bi prišlo do velikega pritiska na vlade držav v tej smeri. Vendar pa ni videti, da bi bile vlade v zvezi s tem trenutno pod kakršnim koli pritiskom. Nasprotno, zdi se, da bo UI prevzela nadzor (in oblast) na povsem benigen način ob večinskem tihem odobravanju velike večine človeštva. To je scenarij, ki se že uresničuje. Praktično povsod prevladuje prepričanje, da UI opravlja mnoga dela veliko bolje kot ljudje. To je običajno formulirano tako: »Z uporabo UI smo pri reševanju mnogih nalog vsi veliko učinkovitejši, zato bi bilo nespametno, da pri vseh teh nalogah ne bi uporabljali UI.« Ta praksa, ko ljudje sami od sebe z racionalnim razmislekom o sposobnosti UI tej prepuščajo vse več pomembnega dela in pomembnih odločitev, se hitro uveljavlja na pomembnih področjih. Ta praksa velja v vse večji meri npr. tudi za področje pravosodja – na primer pisanje obtožnih predlogov, pripravljalnih vlog in sodb. Po vsej logiki bo to kmalu veljalo, ali pa že velja, tudi za pisanje zakonov. Po enaki logiki bo tudi za presojo predlogov zakonov najbolj kompetentna UI, in neracionalno bi bilo, da bi parlamentarna telesa in poslanci sami ocenjevali zakone, ki jih napiše UI. Najkompetentneje jih bo ocenila UI. S tem bo človeštvo samo prepustilo nadzor nad zakonodajo umetni inteligenci. Lahko, da bo taka zakonodaja boljša in učinkovitejša, vendar je vprašanje, ali si to res želimo, ali to ne vodi do odtujitve človeku? Po podobni logiki bo to veljalo za področja znanosti, umetnosti in izobraževanja. Mnogi menijo, da je to pot brez možnosti povratka.

Zahvala

Prispevek je deloma nastal v okviru ciljnega raziskovalnega projekta V2-2272 Opredelitev okvira za zagotavljanje zaupanja javnosti v sisteme umetne inteligence in njihove uporabe, ob podpori Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije in Ministrstva za digitalno preobrazbo Republike Slovenije.

Citirana literatura

Ai, Lun, Johannes Langer, Stephen H. Muggleton, in Ute Schmid. 2023. »Explanatory Machine Learning for Sequential Human Teaching«. *Machine Learning Journal* 112: 3591–3632. https://doi.org/10.1007/s10994-023-06351-8.

Alelyani, Salem. 2021. »Detection and Evaluation of Machine Learning Bias«. *Applied Sciences* 11 (14). https://doi.org/10.3390/app11146271.

Alemohammad, Sina, et al. 2023. »Self-Consuming Generative Models Go MAD«. *arXiv*:2307.01850. https://arxiv.org/pdf/2307.01850.

Ali, Sajid, et al. 2023. »Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence«. *Information Fusion*, 99. https://doi.org/10.1016/j.inffus.2023.101805.

Alvarez-Melis, David, in Tommi S. Jaakkola. 2018. »Towards Robust Interpretability with Self-Explaining Neural Networks *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Kanada.

Amgoud, Leila, in Jonathan Ben-Naim. 2022. »Axiomatic Foundations of Explainability«. *Proceedings of the Thirty-First International joint Conference on Artificial Intelligence (IJACAI-22)*: 636–642.

Angwin, Julia , Lauren Kirchner, Jeff Larson in Surya Mattu. 2016. »Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks«. *ProPublica*.

Barredo Arrieta, Alejandro. et al. 2019. »Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI«. arXiv:1910.10045v2. https://doi.org/10.48550/arXiv.1910.10045.

Berk, Richard, et al. 2021. »Fairness in Criminal Justice Risk Assessments: The State of the Art«. *Sociological Methods & Research* 50 (1): 3–44. https://doi.org/10.1177/0049124118782533.

Blanzeisky, William, in Pádraig Cunningham. 2021. »Algorithmic Factors Influencing Bias in Machine Learning«. *arXiv preprint*. https://doi.org/10.48550/arXiv.2104.14014.

Bohanec, Marko, in Ivan Bratko. 1994. »Trading Accuracy for Simplicity in Decision Trees«. *Machine Learning Journal* 15: 223–250. https://doi.org/10.1007/BF00993345.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bratko, Ivan. 1997. »Machine Learning: between Accuracy and Interpretability«. V *Learning, Networks and Statistics*, ur. G. Della Riccia. Dunaj: Springer.

Bratko, Ivan. 2018. »AlphaZero: What's Missing?«. *Informatica* 42 (1).

Bratko, Ivan, in Peter Mulec. 1980. »An Experiment in Automatic Learning of Diagnostic Rules«. *Informatica* 4 (4): 18–25.

Briesch, Martin, Dominik Sobania in Franz Rothlauf. 2023. »Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop«. *arXiv*:2311.16822.

Bubeck, Sebastien, et al. 2023. »Sparks of Artificial General Intelligence: Early Experiments with GPT-4«. *arxiv*:2303.12712v5. https://arxiv.org/abs/2303.12712.

Cestnik, Bojan, Igor Kononenko in Ivan Bratko. 1987. »ASSISTANT 86: A Knowledge-Elicitation Tool for Sophisticated Users«. V *Progress in Machine Learning: Proceedings of European Working Session on Learning EWSL 87*, ur. I. Bratko in N. Lavrač, 31–45. Sigma Press.

Chakraborti, Tathagata , Sarath Sreedharan in Subbarao Kambhampati. 2020. »The Emerging Landscape of Explainable Automated Planning & Decision Making«. *arXiv*:2002.11697. https://doi.org/10.48550/arXiv.2002.11697.

Chen, Zixi, Varshini Subhash, Marton Havasi, Weiwei Pan in Finale Doshi-Velez. 2022. »What Makes a Good Explanation?: A Harmonized View of Properties of Explanations«. *Progress and Challenges in Building Trustworthy Embodies AI (TEA 2022) co-located with NeurIPS 2022.*

Corbett-Davies, Sam, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff in Sharad Goel. 2018. »The Measure and Mismeasure of Fairness«. *Journal of Machine Learning Research* 24 (2023): 1–117. *arXiv preprint.* https://doi.org/10.48550/arXiv.1808.00023.

Courtland, Rachel. 2018. »Bias Detectives: The Researchers Striving to Make Algorithms Fair«. *Nature* 558: 357–360. https://doi.org/10.1038/d41586-018-05469-3.

Dastin, Jeffrey. 2018. »Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women«. *Reuters*, 11. 10. 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Doshi-Velez, Finale, in Been Kim. 2017. »Towards A Rigorous Science of Interpretable Machine Learning«. *arXiv*:1702.08608.

Dressel, Julia, in Hany Farid. 2018. »The Accuracy, Fairness, and Limits of Predicting Recidivism«. *Science Advances* 4 (1).

Edmonds, Mark, et al. 2019. »A Tale of Two Explanations: Enhancing Human Trust by Explaining Robot Behavior«. *Science Robotics* 4 (37): 1–13.

Evropski parlament. 2024. *Artificial Intelligence Act.*

Farič, Ana, in Ivan Bratko. 2023. »Pristranskost v strojnem učenju: dileme in odgovori«. V *Informacijska družba – IS 2023: Kognitivna znanost, Zbornik 26. mednarodne multikonference,* ur. Anka Slana Ozimič, Borut Trpin, Toma Strle in Olga Markič, 10–13. Ljubljana: IJS.

Flores, Anthony W., Kristin Bechtel in Christopher T. Lowenkamp. 2016. »False Positives, False Negatives, and False Analyses: A Rejoinder to ‚Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks'«. *Federal Probation Journal* 80 (2): 38–46.

Fui-Hoon Nah, Fiona, Ruilin Zheng, Jingyuan Cai, Keng Siau in Langtao Chen. 2023. **»**Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration**«**. *Journal of Information Technology Case and Application Research* 25 (3): 277–304. https://doi.org/10.1080/15228053.2023.2233814.

Future of Life Institute Open Letters. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter. https://futureoflife.org/fli-open-letters/

Gordon, Diana F., in Marie Desjardins. 1995. »Evaluation and Selection of Biases in Machine Learning«. *Machine Learning* 20: 5–22. https://doi.org/10.1023/A:1022630017346.

Guidotti, Riccardo, et al. 2018. »A Survey of Methods for Explaining Black Box Models«. *ACM Computing Surveys* 51 (5): 1–42. https://doi.org/10.1145/3236009.

Gunning, David, Eric Vorm, Jennifer Yunyan Wang in Matt Turek. 2021. »DARPA's Explainable AI (XAI) Program: A Retrospective«. *Applied AI Letters* 2 (4). https://doi.org/10.1002/ail2.61.

Hall, Patrick, SriSatish Ambati and Wen Phan. 2017. »Ideas on Interpreting Machine Learning«. *O'Reilly*, 15. 3. 2017. https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/.

Hardt, Moritz, Eric Price in Nathan Srebro. 2016. »Equality of Opportunity in Supervised Learning«. *arXiv preprint*. https://doi.org/10.48550/arXiv.1610.02413.

Hellström, Thomas, Virginia Dignum in Suna Bensch. 2020. »Bias in Machine Learning – What Is It Good for?«. *arXiv preprint*. https://doi.org/10.48550/arXiv.2004.00686.

Holsinger, Alexander M., et al. 2018. »A Rejoinder to Dressel and Farid: New Study Finds Computer Algorithm Is More Accurate than Humans at Predicting Arrest and as Good as a Group of 20 Lay Experts«. *Federal Probation Journal*, 82 (2): 51–56.

Holzinger, Andreas, Anna Saranti, Christoph Molnar, Przemyslaw Biecek in Wojciech Samek. 2022. »Explainable AI Methods – A Brief Overview«. *xxAI* 2020, *LNAI* 13200: 13–38. https://doi.org/10.1007/978-3-031-04083-2_2.

Hüllermeier, Eyke, Thomas Fober in Marco Mernberger. 2013. »Inductive Bias«. *Encyclopedia of Systems Biology*. https://doi.org/10.1007/978-1-4419-9863-7_927.

Hsieh, Cheng-Yu, et al. 2021. »Evaluations and Methods for Explanation through Robustness Analysis«. *arXiv*:2006.00442. https://doi.org/10.48550/arXiv.2006.00442.

Ignatiev, Alexey, Nina Narodytska in Joao Marques-Silva. 2019. »On Validating, Repairing and Refining Heuristic ML Explanations«. *arXiv*:1907.02509. https://doi.org/10.48550/arXiv.1907.02590.

Kleinberg, Jon, Sendhil Mullainathan in Manish Raghavan. 2016. »Inherent Trade-Offs in the Fair Determination of Risk Scores«. *arXiv preprint*. https://doi.org/10.48550/arXiv.1609.05807.

Kolmogorov, A. N. 1963. »On Tables of Random Numbers«. *The Indian Journal of Statistics, Series A* 25: 369–375.

Košmrlj, Lea, in Ivan Bratko. 2025. »Vpliv generativne umetne inteligence na demokratične procese«. V *Od tržne konkurenčnosti tehnologije k družbeni tehnološki zrelosti*, ur. Slavko Splichal, 69–85. Ljubljana: SAZU.

Kurzweil, Ray. 2005. *The Singularity Is Near*. New York: Viking.

LeCun, Yann, Yoshua Bengio in Geoffrey Hinton. 2015. »Deep Learning«. *Nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Mehrabi, A., F. Morstatter, N. Saxena, K. Lerman in A. Galstyan. 2021. *ACM Computing Surveys* (CSUR) 54 (6): 1–35. https://doi.org/10.48550/arXiv.1908.09635.

Michie, Donald. 1988. »Machine Learning in the Next Five Years«. *Proceedings of the 3rd European Working Session on Learning*: 107–122. https://dl.acm.org/doi/10.5555/3108771.3108781.

Mirzadeh, Iman, et al. 2024. »GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models«. *arXiv*:2410.05229v1. https://arxiv.org/abs/2410.05229.

Mohseni, Sina, Jeremy E Block in Eric Ragan 2021. »Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark«. *2*6th International Conference on Intelligent User Interfaces (IUI'21). https://doi.org/10.1145/3397481.3450689.

Muggleton, Stephen, Ute Schmid, Christina Zeller, Alireza Tamaddoni-Nezhad in Tarek Besold. 2018. »Ultra-Strong Machine Learning: Comprehensibility of Programs Learned with ILPP«. *Machine Learning* 107: 1119–1140. https://link.springer.com/article/10.1007/s10994-018-5707-3.

Ntoutsi, Eirini, et al. 2020. »Bias in Data-Driven Artificial Intelligence Systems – An Introductory Survey«. *WIREs Data Mining and Knowledge Discovery* 10 (3). https://doi.org/10.1002/widm.1356.

Plaat, Aske, et al. 2024. »Reasoning with Large Language Models, a Survey«. *arXiv*:2407.11511. https://arxiv.org/abs/2407.11511.

Poole, David L., in Alan K. Mackworth. 2023. *Artificial Intelligence: Foundations of*

Computational Agents. Tretja izdaja. Cambridge University Press.

Quinlan, J. Ross. 1979. *Discovering Rules by Induction From Large Collections of Examples*. *Expert Systems in the Micro Electronics Age*. Edinburgh University Press.

Ribeiro, Marco Tulio, Sameer Singh in Carlos Guestrin. 2016. »'Why Should I Trust You?': Explaining the Predictions of Any Classifier«. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778.

Ribera, Mireia, in Agata Lapedriza. 2019. »Can We Do Better Explanations? A Proposal of User-Centered Explainable AI«. *CEUR Workshop Proceedings, 2327. https://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf.*

Rudin, Cynthia. 2019. »Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead**«**. *Nature Machine Intelligence* 1 (5): 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Russell, Stuart, in Peter Norvig. 2020. *Artificial Intellignece: A Modern Approach*. **Četrta** izdaja. Harlow: Pearson Education.

Spielkamp, Matthias. 2017. **»**Inspecting Algorithms for Bias**«**. *MIT Technology Review.* https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/.

Sun, Wenlong, Olfa Nasraoui in Patrick Shafto. 2020. »Evolution and Impact of Bias in Human and Machine Learning Algorithms Interaction«. *PLoS ONE* 15 (18). https://doi.org/10.1371/journal.pone.0235502.

Šoberl, Domen, in Ivan Bratko. 2023. »Transferring a Learned Qualitative Cart-Pole Control Model to Uneven Terrains«. International Conference on Discovery Science, 446–459. http://dx.doi.org/10.1007/978-3-031-45275-8_30.

Šoberl, Domen, in Ivan Bratko. 2025. »Qualitative Control Learning Can Be Much Faster than Reinforcement Learning«. *Machine Learning Journal* 114: 4. https://doi.org/10.1007/s10994-024-06724-7

UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

van Eecke, P. 2025. Osebna komunikacija.

Yeh, Chih-Kuan, in Pradeep Ravikumar. 2021. »Objective Criteria for Explanations of Machine Learning Models«. *Applied AI Letters* 2 (4) / e 57. https://doi.org/10.1002/ail2.57.

Yu, Han, et al. 2018. »Building Ethics into Artificial Intelligence«. *arXiv preprint.* https://doi.org/10.48550/arXiv.1812.02953.

Zhang, Zheng, Liangliang Xu, Levent Yilmaz in Bo Liu. 2021. »A Critical Review of Inductive Logic Programming Techniques for Explainable AI«. *arXiv*:2112.15319. https://doi.org/10.48550/arXiv.2112.15319.

Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva in Antonio Torralba. 2016. »Learning Deep Features for Disciminative Localization«. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 27.-30. junij 2016, 2921-2929. https://doi.org/10.1109/CVPR.2016.319.