TRANSPARENTNOST IN RAZLOŽLJIVOST KOT ZAHTEVI ZA ZAUPANJA VREDNO UMETNO INTELIGENCO

OLGA MARKIČ
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
olga.markic@ff.uni-lj.si

Sinopsis Umetna inteligenca je dandanes prisotna tako v vsakdanjem življenju kot na različnih področjih znanosti ter družbenega in gospodarskega življenja. Drugi val umetne inteligence se osredotoča na izdelovanje pametnih orodij, ki temeljijo na strojnem učenju. Kljub relativni uspešnosti pri napovedovanju sistemi drugega vala izkazujejo pomanjkljivosti, na katere ob vedno bolj množični uporabi opozarjajo tako računalničarji kot družboslovci in humanisti. Načrtovalci modelov se pogosto ne zavedajo dovolj, da tako učni primeri kot zastavitve ciljev odražajo družbene vrednote in so vpeti v družbeni kontekst. V prispevku bom predstavila Etične smernice za zaupanja vredno umetno inteligenco. Osredotočila se bom predvsem na zahtevi po transparentnosti in razložljivosti, ki sta dve od zahtev za zaupanja vredno umetno inteligenco. Ker modeli strojnega učenja večinoma ne temeljijo na človeku razumljivem logičnem sklepanju, se bom vprašala, v kolikšni meri jih današnji sistemi dejansko lahko izpolnjujejo.

Ključne besede:
umetna inteligenca,
strojno učenje,
razložljivost,
transparentnost,
etične vrednote

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.3
ISBN 978-961-286-675-4

TRANSPARENCY AND EXPLICABILITY AS REQUIREMENTS FOR TRUSTWORTHY AI

OLGA MARKIČ
University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
olga.markic@ff.uni-lj.si

Abstract Artificial intelligence is present everywhere, in everyday life, in science, and in social institutions and economy. The second-wave AI focuses on developing smart tools based on machine learning. Despite relatively good predictive powers, the systems of the second-wave AI have some drawbacks pointed out by computer scientists as well as by social scientists and humanists. Designers of the models too often ignore the fact that training examples and goals reflect social values and are embedded in the social context. I will first present Ethics guidelines for Trustworthy AI. I will focus on the requirements of transparency and explicability, the two requirements necessary for a Trustworthy AI. And, since machine learning models are not based on understandable logical reasoning systems, I will examine whether these two requirements can be fulfilled by today's systems.

Keywords: artificial intelligence, machine learning, explicability, transparency, ethical values

https://doi.org/10.18690/um.ff.11.2022.3 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Umetno inteligenco dandanes uporabljamo tako običajni ljudje v vsakdanjem življenju, ko si, na primer, pomagamo s spletnimi iskalniki, uporabljamo satelitsko navigacijo ali pa nakupujemo po spletu, kot tudi strokovnjaki in znanstveniki na različnih področjih. Računalniški programi, ki temeljijo na matematičnih algoritmih, so bili v preteklosti deležni precejšnjega zaupanja. Razlogi za to so bili verjetno vezani na zaupanje v matematiko, ki je kot deduktivna znanost zagotavljala gotovost spoznanja. Če so izhodiščne premise resnične, ni mogoče, da bi prišli do neresničnega sklepa. V prvem valu umetne inteligence,[1] ki ga pogosto imenujemo kar klasična umetna inteligenca, so se raziskovalci ukvarjali z idejo podvajanja človeških zmožnosti. Osredotočali so se predvsem na psihološko raven, na modeliranje jezika in logično sklepanje, pri čemer so se opirali na deduktivno logiko. Čeprav so bila pričakovanja in napovedi v začetku zelo optimistične, je prvi val zašel v slepo ulico, ki jo označujejo tudi kot »zimo umetne inteligence« (Russell in Norvig 2010). Drugi val, ki smo mu priča zdaj, pa črpa predvsem iz kibernetike s sredine prejšnjega stoletja in modeliranja z nevronskimi mrežami. Poudarek je na indukciji, na učenju na osnovi predhodnih izkušenj in interakciji z okoljem. Posledično se s tem spremeni tudi pristop k oblikovanju modelov. Če je bila za izdelavo sistemov/modelov prvega vala potrebna predhodna analiza procesa reševanja naloge, na podlagi katere je bil potem programiran model, v drugem valu računalničarji oblikujejo algoritme za strojno učenje, pri čemer uporabljajo učne primere oziroma množice podatkov. Drugi val tako zaznamuje razvoj teorij strojnega učenja in povezovanje z drugimi disciplinami, predvsem s statistiko in teorijo verjetnosti. Zaradi velikih baz podatkov, ki jih omogoča internet, se je začel fokus premikati z algoritmov na same podatke.

Vedno bolj razširjena uporaba sistemov umetne inteligence in premik od sistemov prvega vala k sistemom drugega vala pa poleg dobrobiti, ki jih je razvoj nedvomno prinesel, vzbuja tudi nemalo pomislekov in zaskrbljenosti. Izpostavlja se vprašanje zaupanja v te nove tehnološke pristope, na kar opozarjajo predvsem družboslovci in humanisti, težav pa se začenjajo zavedati tudi računalničarji. Po eni strani gre za bojazen, da zaradi interesov tistih, ki s sistemi upravljajo (podjetja ali država), prihaja do manipulacije uporabnikov. To sicer ni nov pojav, saj željam močnejših, da bi nadzirali in manipulirali, lahko sledimo skozi zgodovino. Dobro poznane so, na primer, manipulacije s pomočjo klasičnih medijev (tisk, radio, televizija), zdaj pa smo priča

[1] Razdelitev na prvi in drugi val je navdahnjena z Cantwell Smith (2019) in povzeta po Markič (2021).

manipulacijam s pomočjo družabnih omrežij. Vendar, kot bom pokazala v nadaljevanju, današnja 'pametna' orodja sprožajo vprašanje zaupanja tudi v kontekstih, kjer ni prisotno namerno zavajanje. Sistemi drugega vala, ki uporabljajo strojno učenje, za razliko od orodij klasične umetne inteligence, ne temeljijo na človeku razumljivem logičnem sklepanju. Cilj sistemov globokih nevronskih mrež je prepoznati vzorce, klasificirati in poiskati napovedi. Vendar do rešitev za različne naloge strojnega učenja sistem prihaja na način, ki je potencialno netransparenten[2] za človeka, tako na strani raziskovalca kot na strani uporabnika. Sistem predstavlja nekakšno 'črno škatlo', kjer uporabniki (pogosto pa tudi načrtovalci) nimajo razlage, na kakšen način je sistem prišel do končnega rezultata, kar zmanjšuje zaupanje. Zato se je pojavila potreba po urejanju področja na način, ki bo skladen s sicer sprejetimi demokratičnimi standardi. V prispevku bom predstavila Etične smernice za zaupanja vredno umetno inteligenco, temeljne vrednote in zahteve, ki bi jih moral sistem umetne inteligence izpolniti, ter se osredotočila na vprašanje razložljivosti in transparentnosti.

2 Umetna inteligenca in algoritmi

Kot je bilo omenjeno v uvodu, se je opredelitev umetne inteligence spreminjala skozi čas. V filozofskih diskusijah (glej npr. Bringsjord in Govindarajulu 2020; Markič 2021) se največkrat deli na splošno oziroma močno umetno inteligenco, katere končni cilj je ustvariti »stroje, ki mislijo, se učijo in ustvarjajo« (Simon v Russell in Norvig 2010: 27), in šibko oziroma ozko umetno inteligenco, ki razvija predvsem pametna orodja na izbranem področju. V tem prispevku se bomo osredotočili na slednjo, saj v tem trenutku najbolj zaznamuje naše življenje. V dokumentu Organizacije za gospodarsko sodelovanje in razvoj (OECD) je umetna inteligenca opredeljena takole:

> Sistem umetne inteligence je strojni sistem, ki lahko vpliva na okolje s podajanjem napovedi, priporočil ali odločitev za dano množico podatkov. Uporablja podatke, pridobljene strojno in/ali s pomočjo človeka, zato da (i) zaznava resnično in/ali virtualno okolje; (ii) abstrahira te zaznave v modele s pomočjo avtomatske analize (npr. strojnim učenjem) ali ročno; (iii) s pomočjo modelov sklepanja predvidi možne izide. Sistemi umetne inteligence so

[2] Sama raje uporabljam izraza transparentnost, (ne)transparenten, v slovenskih tekstih sta sinonimno uporabljena tudi izraz preglednost, (ne)pregleden (npr. v Smernicah za zaupanja vredno umetno inteligenco 2019).

> oblikovani tako, da delujejo z različnimi stopnjami avtonomije. (OECD.AI Policy Observatory n.d.)

Če je bila umetna inteligenca v začetkih predvsem domena znanstvenikov, v javnost pa je prišla preko filmov in znanstvene fantastike, pa je, kot smo omenili že v uvodu, dandanes prisotna že v našem vsakdanjem življenju. Ed Finn (2017) poudarja, da se je naš odnos do računalnikov spremenil proti koncu prvega desetletja našega stoletja, ko smo v žepih kot zveste spremljevalce začeli nositi pametne telefone in namesto o strojni opremi začeli govoriti o aplikacijah in uslugah. Telefoni niso bili več samo pripomočki, ki jih občasno uporabljamo, ampak smo jim začeli zaupati pri izbiri poti, prijateljev in vsebin, vrednih ogleda. Kot pravi Finn, smo z vsakim klikom in sprejemom pogojev uporabe sprejeli idejo, da veliki podatki, senzorji in različne oblike strojnega učenja lahko modelirajo in uravnavajo vse vrste kompleksnih sistemov, od izbire pesmi do napovedi kriminala. Ključno vlogo pa je prevzel izraz algoritem. Algoritmi so povsod, prevladujejo na borzah, skladajo glasbo, vozijo avtomobile, pišejo članke. (Finn 2017: 15) Sama beseda algoritem naj bi izhajala iz latiniziranega imena Algoritmi za perzijskega matematika iz 9. stoletja al-Khwārizma (Gillespie 2016: 19). Skozi čas je izraz algoritem dobil pomen postopka, ki opisuje množico matematičnih navodil za manipuliranje podatkov oziroma postopek, ki kar najhitreje pripelje do želenega rezultata. Na primer, Evklidov algoritem za iskanje največjega skupnega delitelja. Finn poudarja, da se je v računalništvu uveljavila pragmatistična opredelitev – kot »metoda za reševanje problema«, »osvetljevanje poti med problemi in rešitvami« (Finn 2017: 18). Donald Knuth je v klasičnem delu *The Art of Computer Programming* zapisal: »algoritem je končna množica pravil, ki daje zaporedje za reševanje določenega tipa problemov« (Knuth 1997: 4). A kot opozarja Terleton Gillespie, je algoritem zgolj recept, ki ga sestavljajo programabilni koraki. Pred tem mora biti opredeljen model, ki dejansko formalizira problem, opredeli cilj in ga predstavi v računskih izrazih. Kompleksna družbena aktivnost in vrednote so prevedene v funkcionalne interakcije spremenljivk in korakov. Vprašanje, kaj je relevantno, kaj pa je družbena sodba, postane del modela. (Gillespie 2016: 19–20). Različni algoritmi lahko znotraj danega modela dosežejo isti rezultat. Na primer, različni algoritmi za razvrščanje po abecedi, ki pa se lahko razlikujejo po hitrosti. A tisto, kar predstavlja družbeno relevantno razliko, se bolj nanaša na sam problem, ki ga rešujemo, na način, kako predstavimo spremenljivke in kako izberemo ter predstavimo cilj. V našem primeru bi se lahko vprašali, zakaj sploh razvrščati po abecedi.

V širši javnosti in tudi med družboslovnimi raziskovalci je izraz algoritem prevzel mnogo širši pomen. Kot poudarja Gillespie, je postal okrajšava za vse prej povedano: za algoritem v ožjem smislu, model, izbrani cilj, podatke, učenje na podatkih, aplikacijo, strojno opremo. Tako je postal ime za določeno vrsto družbenotehničnega sistema, ki proizvaja znanje in se odloča, in v katerem so ljudje, reprezentacije in informacije ponujeni kot podatki, ki so eden z drugim postavljeni v sistematične/matematične odnose, in kjer jim je pripisana izračunana vrednost. Gre za besedno figuro sinekdoho, ko se celota poimenuje z njenim delom. (Gillespie 2016: 22–23) Če to spregledamo, ne bomo dobro razumeli, zakaj se ljudje jezijo nad Facebookovimi in Googlovimi algoritmi. Razumljeni v natančnem ozkem pomenu so algoritmi samo zaporedje navodil, a kar ljudi moti, je razumevanje v širšem smislu, kjer v procesu igrajo pomembno vlogo tudi vrednote in človeške odločitve (npr. kaj izberemo za cilj – je to zasledovanje čim večjega dobička ne glede na negativne posledice za uporabnike in družbo, kot je izpostavila Frances Haugen v svoji kritiki Facebooka). Dejansko niti ni pomembno, ali celoto imenujemo model, sistem ali algoritem, če se zavedamo, na kaj se zanašamo. A ker ima izraz algoritem dva pomena, lahko prihaja do ekvivokacije. Na primer, ponudnik poudarja, da aplikacija temelji na preverjenem algoritmu, pri čemer sugerirajo ozek pomen, nič pa na primer ne povedo, kako so bili predstavljeni in izbrani podatki, kar je dejansko pomembno pri vrednotenju aplikacije. Uporabniki tako pomislijo na šolska leta in na natančnost matematičnih algoritmov ter zato hitreje in brez pomislekov sprejmejo aplikacijo kot zaupanja vredno. A kot bom pokazala v nadaljevanju, je dandanes veliko algoritmov (razumljenih v širšem smislu) oziroma sistemov umetne inteligence, ki so netransparentni, kar uporabniku, pogosto pa tudi načrtovalcu, onemogoča razlago in razumevanje delovanja. Vprašanje, katere so pomembne sestavine sistema in na kaj moramo biti pozorni, ko jih vrednotimo, je toliko bolj pomembno, ker algoritmi postajajo vse bolj prisotni v našem življenju. Kot poudarja Danaher, gre za »neizogibno in vseprisotno uporabo računalniških algoritmov za razumevanje in nadzor sveta, v katerem živimo« (Danaher 2020: 2). Zato nekateri govorijo kar o vladavini algoritmov oziroma o »algokraciji« (Aneesh 2006; Danaher 2016; 2020).[3]

Sistemi umetne inteligence so torej predvsem orodja, ki naj bi pomagala človeku, a hkrati tudi ključno posegajo v družbene odnose in v intimo ljudi. Zato ta orodja ne morejo biti samo stvar inženirjev in znanstvenikov s področja tehnike, ampak se tičejo vseh uporabnikov. V zadnjih letih smo bili priča odmevnim nastopom

[3] Več o tem v poglavju »Umetna inteligenca, algokracija in avtonomija« (Strle in Markič 2021).

žvižgačic in žvižgačev, ki so opozorili, da velike korporacije zaradi želje po dobičku kršijo celo lastna pravila in etična načela. Odmevno razkritje leta 2018 je bila zloraba več deset milijonov Facebook računov Američanov za namene vplivanja na volitvah s strani Cambridge Analytica, svetovalne agencije, ki je delovala za Donalda Trumpa v volilni kampanji leta 2016. Pojavili so se strahovi, da algoritmi, ki določajo, kaj ljudje vidijo na platformi, dejansko povečujejo lažne novice in sovražni govor, ki jih ruski hekerji uporabljajo za poskus vpliva na volitve v Trumpovo korist (Hao 2021a). Da bi si povrnili ugled, so pri Facebooku osnovali skupino z imenom Odgovorna UI (angl. *Responsible AI*), a dejansko so se, predvsem v državah izven območja Severne Amerike in Evrope (Myanmar, Honduras, Etiopija, India), nadaljevale zlorabe lažnih profilov v politične namene. O tem je 2020 zelo glasno spregovorila žvižgačica Sophie Zang (Hao 2021b). O neukrepanju, čeprav so vedeli za težave mladostnikov, ki so z uporabo Instagrama,[4] ki poudarja medvrstniško primerjavo teles in stila življenja, zahajali v velike stiske, je govorila še ena bivša uslužbenka Facebooka, Frances Haugen. Sama se je zato zavzela za nujno zunanjo regulacijo na področju družbenih medijev (Waterson in Milmo 2021). To je le nekaj najodmevnejših sporočil, ki so v običajnih ljudeh vzbudila nezaupanje v algoritme umetne inteligence na družbenih omrežjih. Temu lahko dodamo tudi vedno več kritik raziskovalk in raziskovalcev z akademskega področja, ki opozarjajo na pristranosti, ki vodijo v neetične odločitve (O'Neil 2016; Eubanks 2017) in na pasti s tehnologijo omogočenega nadzorovanja (Zuboff 2019). Jasno postaja, da uporaba sistemov umetne inteligence zahteva širši družbeni in etiški premislek.

3 Etične smernice za zaupanja vredno umetno inteligenco

Prav zaradi kritik in porajajočega se nezaupanja v sisteme umetne inteligence se je pokazalo, da je treba področje začeti urejati tako, da bo skladno z vrednotami demokratične družbe. Ti pritiski se pojavljajo predvsem v zahodnem svetu, so pa zahteve po bolj pravičnem urejanju prisotne po celem svetu.[5] V prispevku bom kot primer izpostavila *Etične smernice za zaupanja vredno umetno inteligenco*,[6] ki jih je kot priporočilo izdala Strokovna skupina na visoki ravni za umetno inteligenco pri Evropski komisiji leta 2019. Kot ugotavljajo, je treba priznati in upoštevati:

[4] V lasti Facbooka, zdaj Mete.
[5] V Sloveniji na primer deluje mednarodni center za umetno inteligenco IRCAI (International Center for Artificial Intelligence) pod okriljem Unesca (https://ircai.org/).
[6] V nadaljevanju *Etične smernice*.

> da sistemi umetne inteligence posameznikom in družbi sicer prinašajo znatne koristi, vendar predstavljajo tudi nekatera tveganja in imajo lahko negativne vplive, tudi take, ki jih je morda težko predvideti, opredeliti ali izmeriti (npr. vpliv na demokracijo, pravno državo in pravično porazdelitev ali na sam človeški um). Po potrebi je treba sorazmerno z obsegom tveganja sprejeti ustrezne ukrepe za zmanjšanje teh tveganj. (Etične smernice 2019: 2) .

V Smernicah so izpostavili tri elemente:

1. morala bi biti zakonita ter spoštovati vse veljavne zakone in druge predpise,
2. morala bi biti etična ter zagotavljati spoštovanje etičnih načel in vrednot ter
3. morala bi biti robustna s tehničnega in družbenega vidika, saj lahko sistemi umetne inteligence povzročijo nenamerno škodo, tudi če se uporabljajo z dobrimi nameni. (Etične smernice 2019: 6)

Temelje zaupanja vredne umetne inteligence predstavljajo štiri etična načela oziroma zahteve, ki temeljijo na temeljnih pravicah: spoštovanje človekovega dostojanstva; svoboda posameznika; spoštovanju demokracije, pravičnosti in pravne države; enakost, nediskriminacija in solidarnost; pravice državljanov (Etične smernice 2019: 12–13).

Ta štiri načela so:

(i) spoštovanja človekove avtonomije,
(ii) preprečevanja škode,
(iii) pravičnosti,
(iv) razložljivosti. (Etične smernice 2019: 14)

Izpolnjevanje prvega načela od načrtovalcev sistemov umetne inteligence zahteva, da spoštujejo temeljne pravice, na katerih temelji EU in so namenjene zagotavljanju spoštovanja svobode in avtonomije ljudi.

> Ljudje, ki komunicirajo s sistemi umetne inteligence, morajo imeti možnost, da se še naprej v celoti in dejansko samostojno odločajo in sodelujejo v demokratičnem procesu. Sistemi umetne inteligence si ne bi smeli neupravičeno podrediti ljudi ali jih siliti, zavajati, manipulirati z njimi, jih določati ali zbirati v skupine. (Etične smernice 2019: 14)

Razprave o tem, ali v kakšni meri to zahtevo izpolnjujejo aktualni sistemi umetne inteligence, kje so nevarnosti in v kakšno smer bi moral iti bodoči razvoj, so dandanes vroča tema (več o tem glej Strle in Markič 2021). Prej omenjene manipulacije, ki smo jim priča uporabniki, nas opozarjajo, da je to načelo pogosto kršeno. Včasih tudi z izgovorom, da tako narekuje izpolnjevanje drugega načela, ki govori o preprečevanju škode. Na primer, zaradi varnosti in preprečevanja kriminala in terorističnih napadov se uporablja nadzor (npr. kamere s prepoznavanjem obrazov), ki posega v avtonomijo človeka. Zavedati se je treba, da so trenja med posameznimi etičnimi načeli, ki so zapisna v abstraktni obliki, etične dileme, ki nimajo enostavnih tehnoloških rešitev, temveč zahtevajo poglobljen razmislek. A kot je zapisano, so »nekatere temeljne pravice in soodvisna načela [so] absolutni in se ne bi smeli tehtati (npr. človekovo dostojanstvo)« (Etične smernice 2019: 16). Prav zato je EU tudi predlagala nova pravila o uporabi na področjih, kjer gre za nesprejemljivo tveganje. »Vse, kar se šteje za očitno grožnjo za državljane EU, bo prepovedano: od 'družbenega točkovanja' vlad do igrač z glasovnim upravljanjem, ki spodbujajo k nevarnemu ravnanju otrok« (Evropska komisija n.d.).

Ta načela bi morala biti nato preoblikovana v konkretne zahteve za doseganje zaupanja vredne umetne inteligence in bi morala veljati za različne deležnike, ki sodelujejo v življenjskem ciklu sistemov umetne inteligence: razvijalce, uvajalce in končne uporabnike ter širšo družbo (Etične smernice 2019: 16). V tem prispevku se bom v razpravi o zaupanja vredni umetni inteligenci omejila predvsem na zadnje načelo oziroma zahtevo po razložljivosti in njeno navezavo na postopkovno pravičnost ter na bolj konkretne zahteve po preglednosti.

Načelo razložljivosti in postopkovna pravičnost sta v *Etičnih smernicah* opredeljeni takole:

> Razložljivost je ključna za vzpostavljanje in ohranjanje zaupanja uporabnikov v sisteme umetne inteligence. To pomeni, da morajo biti postopki pregledni, da je treba odkrito sporočati zmogljivosti in namen sistemov umetne inteligence

> ter da mora biti mogoče odločitve – kolikor je mogoče – razložiti tistim, na katere neposredno in posredno vplivajo. Brez takšnih informacij odločitev ni mogoče ustrezno izpodbijati. Ni vedno mogoče pojasniti, zakaj je model dal določen rezultat ali odločitev (in katera kombinacija vhodnih dejavnikov je prispevala k temu). Ti primeri se imenujejo algoritmi „črne škatle" in jim je treba nameniti posebno pozornost. V navedenih okoliščinah je treba morda sprejeti druge ukrepe za razložljivost (npr. sledljivost, možnost revidiranja in pregledno obveščanje o zmogljivostih sistema), če sistem kot celota spoštuje temeljne pravice. Potrebna stopnja razložljivosti je zelo odvisna od okoliščin in resnosti posledic, če je navedeni rezultat napačen ali drugače netočen[7]. (Etične smernice 2019: 15)

> Postopkovna razsežnost pravičnosti vključuje zmožnost izpodbijanja odločitev, ki jih sprejmejo sistemi umetne inteligence in ljudje, ki jih upravljajo, ter zmožnost uveljavljanja učinkovitega pravnega sredstva proti njim. V ta namen mora biti mogoče identificirati subjekt, ki je odgovoren za odločitev, postopki odločanja pa bi morali biti razložljivi. (Etične smernice 2019: 15)

Načelo razložljivosti podpira bolj konkretna zahteva po preglednosti elementov, pomembnih za sistem umetne inteligence: podatkov, sistema in poslovnih modelov. Vključuje sledljivost, razložljivost in obveščanje.

> **Sledljivost**. Nabore podatkov in procese, na podlagi katerih se s sistemom umetne inteligence sprejme odločitev, vključno s tistimi o zbiranju in označevanju podatkov, ter uporabljene algoritme bi bilo treba dokumentirati v skladu z najboljšimi možnimi standardi, da se omogočita sledljivost in povečanje preglednosti. To velja tudi za odločitve, ki jih sprejme sistem umetne inteligence. To omogoča opredelitev razlogov, zakaj je bila odločitev umetne inteligence napačna, kar pa bi lahko pomagalo preprečiti prihodnje napake. Zato sledljivost omogoča revidiranje in razložljivost.

[7] Na primer netočna priporočila sistema umetne inteligence pri nakupovanju vzbujajo manj pomembne etične pomisleke kot sistemi umetne inteligence, ki ocenjujejo, ali naj se posameznika, obsojenega za kaznivo dejanje, pogojno izpusti.

> **Razložljivost.** Razložljivost se nanaša na zmožnost pojasniti tehnične procese sistema umetne inteligence in s tem povezane človeške odločitve (npr. področja uporabe sistema umetne inteligence). V skladu s tehnično razložljivostjo morajo ljudje razumeti odločitve, ki jih sprejme sistem umetne inteligence, in jim biti zmožni slediti. Poleg tega je morda treba doseči kompromise med povečanjem razložljivosti sistema (ki lahko zmanjša njegovo točnost) ali povečanjem njegove točnosti (na račun razložljivosti). Kadar sistem umetne inteligence pomembno vpliva na življenje ljudi, bi moralo biti mogoče zahtevati ustrezno razlago postopka odločanja sistema umetne inteligence. Takšna razlaga bi morala biti pravočasna in prilagojena strokovnemu znanju zadevnega deležnika (npr. nestrokovnjak, regulator ali raziskovalec). Poleg tega bi morala biti na voljo pojasnila o tem, koliko sistem umetne inteligence vpliva na organizacijski postopek odločanja in ga oblikuje, in pojasnila o izbiri zasnove sistema ter utemeljitev za njegovo uvedbo (kar bi zagotovilo preglednost poslovnega modela).
>
> **Obveščanje**. Sistemi umetne inteligence se ne bi smeli uporabnikom predstavljati kot ljudje; ljudje imajo pravico do seznanitve s tem, da so v stiku s sistemom umetne inteligence. To pomeni, da morajo biti sistemi umetne inteligence kot taki prepoznavni. Poleg tega bi bilo treba za zagotovitev skladnosti s temeljnimi pravicami po potrebi zagotoviti možnost, da se uporabniki odločijo za komuniciranje s človekom namesto s sistemom umetne inteligence. Nadalje, strokovnjake na področju umetne inteligence ali končne uporabnike bi bilo treba obveščati o zmožnostih in omejitvah sistema umetne inteligence, in sicer na način, primeren za zadevni primer uporabe. To bi lahko zajemalo obveščanje o stopnji točnosti sistema umetne inteligence in njegovih omejitvah. (Etične smernice 2019: 21–22)

Etične smernice izpostavljajo načela in na njih temelječe zahteve za zaupanja vredno umetno inteligenco. Vprašanje je, kako deležnike zavezati k njihovem spoštovanju, če to ni v njihovem interesu (npr. velike korporacije, kot so Google, Meta, Amazon, Tik Tok itd.). Na tem področju čaka družbo, predvsem pravnike, še veliko dela.

V zadnjem razdelku bom skušala pokazati, zakaj, četudi bi iskreno želeli spoštovati načelo razložljivosti in zahtevo po transparentnosti, to ni enostavno in je morda v obliki, ki bi zadostila standardom, kot jih želimo na področju prava in medicine, trenutno tudi neizvedljivo.

4 Sistemi strojnega učenja in zahtevi po razložljivosti in transparentnosti

Kot smo omenili v uvodu, je drugi val umetne inteligence zasnovan na induktivnih oblikah sklepanja – računalničarji namesto algoritmov kot navodil (v obliki pravil), kako rešiti določeno nalogo, pišejo algoritme za strojno učenje. Pomembno vlogo pri tem pristopu igrajo podatki, na katerih se sistem uči oziroma iz katerih sistem razbere vzorce, ki pomagajo pri napovedih (rudarjenje podatkov – angl. *data mining*). Sistemi, ki so v zadnjih letih področje umetne inteligence približali uporabnikom, temeljijo na različnih pristopih strojnega učenja. Dejansko se je začel razvoj takih modelov pospešeno razvijati v 80-ih letih prejšnjega stoletja, ko so računalničarji z odkritjem posplošenega delta pravila (angl. *back propagation rule*) lahko učili tudi večnivojske mreže (Rumelhart et al. 1986). Kot sva s kolegom Strletom na kratko predstavila v razdelku *Umetna inteligenca: prvi in drugi val* (Strle in Markič 2021), bi strojno učenje v grobem lahko razdelili na nadzorovano učenje, kjer se sistem uči na podlagi učnih primerkov in poznanih rezultatov, na nenadzorovano učenje, kjer se sistem uči sam v interakciji z okoljem, in na spodbujevano učenje, kjer se sistem v daljšem obdobju prosto odloča, ob vsaki odločitvi pa prejme nagrado, če je bila odločitev dobra, oziroma kazen, če je bila slaba. Taki sistemi so se zgledovali po delovanju živčnih mrež v možganih, zato jih pogosto imenujemo nevronske mreže. Vendar nas ime ne sme zavesti. Cilj sodobnih sistemov umetne inteligence je izdelovanje orodij na različnih področjih življenja (npr. v bančništvu, medicini, športu, pravu in vojski), ki so lahko zelo oddaljena od dejanskega delovanja živčnega sistema[8]. (Strle in Markič 2021: 104–105)

Že kmalu po uveljavitvi nevronskih mrež kot primernih modelov za klasificiranje pa so se pokazale tudi težave, ki pestijo take sisteme. Na primer, sistem se na podlagi učnih primerov nauči prepoznati določen predmet, a nauči se na osnovi zmotnih (angl. *spurious)* korelacij, ne pa na vzročnih povezavah. O takih pomanjkljivostih poročajo tudi pri sodobnih sistemih. Na primer, Lapuschkin in sodelavci (2016) so ugotovili, da pri zmagovalni metodi tekmovanja PASCAL VOC, kjer so bile slike avtomatično pobrane s platforme Flickr, sistem za prepoznavo uporablja korelacije ali kontekst v podatkih. Na primer, čolne prepoznava po prisotnosti vode, vlak po prisotnosti tirnic na sliki. Kar je še bolj šokantno, izkazalo se je, da je sistem

[8] Se pa v računski nevroznanosti uporabljajo orodja drugega vala za znanstveno preučevanje dejanskih nevronskih mrež (živčnega sistema). Npr. Human Brain Project (https://www.humanbrainproject.eu/en/).

prepoznaval konje po zaščitnem vodnem znaku. Kot poudarjata Wojciech Samek in Klaus-Robert Müller (2019), je vodni znak očiten artefakt v zbirki podatkov, ki pa je dolga leta ostal spregledan tako od organizatorjev kot udeležencev tekmovanja. Avtorja tako situacijo primerjata z znanim zgodovinskim primerom 'pametnega Hansa', konja, ki je okoli leta 1900 postal senzacija zaradi domnevne zmožnosti štetja. Kot se je kasneje izkazalo, pa konj ni obvladal matematike, ampak je približno 90 % napovedal pravilen rezultat na osnovi spraševalčevega odziva. Analizirala sta tudi več sodobnih primerov. Recimo, ko se izkaže, da globoka nevronska mreža razločuje med razredoma fotografij 'volk' in 'husky' na podlagi prisotnosti snega. Menita, da napovedovalec, podoben 'pametnemu Hansu', sicer lahko dobro napoveduje na svoji testni zbirki podatkov, a bo odpovedal v realnem svetu, ko čolni niso samo v vodi, ampak se vozijo tudi na prikolicah, ko sta tako volk kot husky lahko oba v okolju brez snega in ko konji nimajo zaščitnega vodnega znaka. A če imamo opraviti s sistemom umetne inteligence, ki predstavlja 'črno škatlo', potem uporabnik zelo težko prepozna, da sistem deluje kot 'pametni Hans' (Samek in Müller 2019: 7–8). Ti in podobni primeri kažejo, da je treba biti mnogo bolj pazljiv pri izbiri podatkov.

Bolj kot take nerodne klasifikacije pa so zanimivi sistemi, ki izkazujejo nekakšno obliko avtonomnega delovanja. Na tem mestu se ne bom spuščala v razpravo, ali bomo računalniškim sistemom kdaj lahko pripisali polno avtonomijo, je pa dejstvo, da se izraz avtonomije uporablja za nekatere sisteme umetne inteligence (npr. avtonomna vozila, avtonomna orožja). Gre za uporabo v šibkejšem pomenu, saj ti sistemi nimajo lastnih intenc in ciljev, te mu še vedno določa človek. Lahko pa takim sistemom pripišemo zmožnost, da sami izberejo pot do cilja. Na primer, znan je primer programa AlphaGo Zero (Silver et al. 2017). Ta se za razliko od svojega predhodnika AlphaGo ni učil iz množice iger mojstrov goja, ampak so ga naučili samo osnovnih pravil postavljanja belih in črnih kamnov, nato pa je, namesto da bi se naslanjal na človeško znanje o igranju goja, prek igre s samim seboj in z metodo spodbujevanega učenja sam odkrival in razvijal svoje 'znanje' o goju. Mindt in Montemayor (2020) poudarjata, da se je AlphaGo Zero na ta način sam naučil igranja, pri čemer je sam tudi odkrival strategije, ki so bile sicer znane mojstrom goja, razvil pa je tudi nekaj novih, ki jih igralci še niso poznali (Mindt in Montemayor 2020: 22–23).

Sistem je tako uspešno dosegel cilj – odlično igrati go in premagovati tekmece, a pot, kako je to dosegel, za človeka ostaja nerazložljiva. Sistem je za človeka kot 'črna škatla', je netransparenten. Ljudje si ne znamo razložiti, na kakšen način je sistem prišel do znanja. V tem konkretnem primeru to morda niti ni tako pomembno (čeprav bi si igralci goja verjetno želeli, da bi dobili razlago, zakaj je določena strategija dobra ali slaba). A sledeč *Smernicam* bi na področjih medicine, prava, zavarovalništva in bančništva ter novinarstva, torej povsod, kjer »sistem umetne inteligence pomembno vpliva na življenje ljudi, moralo biti mogoče zahtevati ustrezno razlago postopka odločanja sistema umetne inteligence« (Etične smernice 2019: 21). Zmožnost, da bi lahko preverili odločitve sistema umetne inteligence, je za zaupanje vanj zelo pomembna tako v situacijah, pri katerih ima sistem podporno vlogo pri naših odločitvah (npr. v medicinski diagnostiki, pri odločanju v pravnem postopku) kot v situacijah, kjer bi sistem praktično sam prevzel odločitve (npr. avtonomna vozila). V kolikor bi sistem umetne inteligence povzročil škodo, mora biti mogoče ugotoviti, zakaj se je to zgodilo. A če do rezultata pridemo tako, da ne razumemo, kaj se dogaja v 'črni škatli', ta zahteva ni izpolnjena.

Poleg tega, da tak sistem 'črne škatle' ne omogoča razlage, kaže še dodatno šibkost. Izkazalo se je, da je globoke nevronske mreže, ki prepoznavajo vzorce, presenetljivo lahko pretentati s slikami, ki so za običajnega človeka vidne kot naključen šum ali abstraktni geometrijski vzorci. Na primer, ko sistem črno rumene črte zmotno prepozna kot šolski avtobus.[9] Kot poroča Davide Castelvecchi v članku s pomenljivim naslovom »Can we open the black box of AI?« in podnaslovom: »Artificial intelligence is everywhere. But before scientists trust it, they first need to understand how machines learn« (2016), pa kljub mnogim predlogom do zdaj ni poznana kakšna splošna rešitev. Nguyen in sodelavci poročajo o tem, kako slike, ki so za ljudi popolnoma neprepoznavne, globoka nevronska mreža z visoko stopnjo verjetnosti klasificira kot znane predmete. Taki rezultati kažejo na zanimive razlike med človeškim vidom in trenutnimi globokimi nevronskimi mrežami in po njihovem mnenju postavljajo vprašanja o splošnosti na globokih nevronskih mrežah temelječega računalniškega vida. (Ngyuen et al. 2015)

[9] V ZDA so šolski avtobusi črno-rumene barve.

V *Etičnih smernicah* je pri zahtevi za razložljivosti opozorilo, da je včasih »morda treba doseči kompromise med povečanjem razložljivosti sistema (ki lahko zmanjša njegovo točnost) ali povečanjem njegove točnosti (na račun razložljivosti)« (Etične Smernice 2019: 21). Ob tem pa je treba upoštevati, da je »potrebna stopnja razložljivosti zelo odvisna od okoliščin in resnosti posledic, če je navedeni rezultat napačen ali drugače netočen« (Etične Smernice 2019: 15). Ta navedek jasno kaže, da *Etične smernice* raziskovalcem puščajo odprto presojo, kakšen kompromis naj napravijo. Menim, da se na tistih področjih, ko gre za pomembne odločitve, ki se tičejo posameznikovega življenja, ne bi smeli odpovedati niti utemeljitvi izbora učnih primerov (podatkov) niti razlagi samega postopka. V kolikor pa se uporabi sistem, ki ne daje človeku razumljive razlage, pa mora strokovnjak - odločevalec to jasno opredeliti in utemeljiti njegovo uporabo in dobljene rezultate v skladu s postopkovno razsežnostjo pravičnosti, kot smo jo navedli v predhodnem razdelku.

5 Zaključek

V prispevku sem pokazala, kako se uporabniki in razvijalci soočajo z vprašanji glede zaupanja v sisteme umetne inteligence. Osredotočila sem se na sisteme drugega vala, ki temeljijo na strojnem učenju in se uporabljajo za napovedovanje, klasificiranje in prepoznavanje vzorcev, zmožni pa so tudi avtonomnega odločanja do neke mere. Vprašanja o zlorabah, manipulacijah in zaupanju v orodja umetne inteligence presegajo zgolj strokovne diskusije znotraj računalništva in so v zadnjem času sprožila val kritičnih odzivov med družboslovci in humanisti, ki opozarjajo predvsem na različne oblike pristranosti in nevarnosti nadzora. Sama sem se v prispevku osredotočila predvsem na epistemski vrednoti transparentnosti in razložljivosti. Sisteme, ki sicer niso transparentni in razložljivi, lahko raziskovalci uspešno uporabljajo v poskusnih (angl. *exploratory*) kontekstih in tako pripomorejo do novih odkritij (Boge 2021). Vendar pa glede na do sedaj dostopne raziskave menim, da na tistih področjih, kjer uporabniki morajo imeti glede odločitve o svojem primeru pravico do razumljive razlage in utemeljitve odločitve, odločanje na podlagi sistemov strojnega učenja ni primerno. Tak sistem za dano nalogo ne bi bil zaupanja vreden.

Viri in literatura

Aneesh, A. (2006). *Virtual Migration*. Durham, NC: Duke University Press.

Boge, F. J. (2021). »Two Dimensions of Opacity and the Deep Learning Predicament«. *Minds and Machines*, 32, str. 43–75.

Bringsjord, S., Govindarajulu, N.S. (2020). »Artificial Intelligence«, V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (izdaja poletje 2020). URL = https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/.

Castelvecchi, D. (2016). »Can we open the black box of AI?«. *Nature*, 538(7623), str. 20–23.

Cantwell Smith, B. (2019). *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA, London: The MIT Press.

Danaher, J. (2016). »The Threat of Algocracy: Reality, Resistance and Accommodation«. *Philosophy and Technology*, 29(3), str. 245–268.

Danaher, J. (2020). »Freedom in an Age of Algocracy«. V Vallor, S. (ur.), *Philosophy of Technology*. Oxford: Oxford University Press, str. 250–272.

European Commission, Directorate-General for Communications Networks, Content and Technology. (2019). *Etične smernice za zaupanja vredno umetno inteligenco*. Publications Office. URL = https://data.europa.eu/doi/10.2759/65329.

Eubaks, V. (2017). *Automating Inequality*. St. Martin's Press: New York.

Evropska komisija. n.d. »Odličnost in zaupanje v umetno inteligenco«. *European Commission* (2. julij 2022). URL = https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_sl#latest.

Finn, E. (2017). *What Algorithms Want: Imagination in the Age of Computing*. Cambridge, MA, London: The MIT Press.

Gillespie, T. (2016). »Algorithm«. V Peters, B. (ur.), *Digital Keywords: A Vocabulary of Information Society and Culture*. Princeton in Oxford: Oxford University Press, str. 18–30.

Hao, K. (2021a). »How Facebook got addicted to spreading misinformation«. *MIT Technology Review* (11. marec 2021). URL = https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

Hao, K. (2021b). »She risked everything to expose Facebook. Now she's telling her story«. *MIT Technology Review* (29. julij 2021). URL = https://www.technologyreview.com/2021/07/29/1030260/facebook-whistleblower-sophie-zhang-global-political-manipulation/.

Knuth, Donald. (1997). *The Art of Computer Programming. Volume 1: Fundamental Algorithms*. Reading: Addison-Wesley.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W. (2016). »Analyzing classifiers: fisher vectors and deep neural networks«. *Conference on Computer Vision and Pattern Recognition (CVPR)*, str. 2912–2920.

Markič, O. (2021). »Prvi in drugi val umetne inteligence«. V Malec, M. in Markič, O. (ur.), *Misli svetlobe in senc: Razprave o filozofskem delu Marka Uršiča*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, str. 201–211.

Mindt, G. in Montemayor, C. (2020). »A Roadmap for Artificial General Intelligence: Intelligence, Knowledge, and Consciousness«. *Mind and Matter*, 18(1), str. 9–37.

Nguyen, A., Yosinski, J. Clune, J. (2015). »Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images«. *Computer Vision and Pattern Recognition (CVPR '15), IEEE*. URL = http://arxiv.org/abs/1412.1897.

OECD.AI Policy Observatory. (n.d.) »OECD AI Principles overview«. *Organisation for Economic Co-operation and Development* (2. julij 2022). URL = https://oecd.ai/en/ai-principles.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens democracy*. Crown: New York.

Russell, S. in Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (*3rd. ed.)*. Upper Saddle River: Prentice Hall.

Samek, W. in Müller, K. R. (2019). »Towards Explainable Artificial Intelligence«. V Samek, W., Montavon, E., Vedaldi A., Hansen, L.K., Müller, K. R. (urd.), *Explainable AI:Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer, str. 5–22.

Silver, D. et al. (2017). »Mastering the Game of Go Without Human Knowledge«. *Nature*, 550(7676), str. 354–359.

Strle, T. in Markič, O. (2021). *O odločanju in avtonomiji*. Maribor: Aristej.

Rumelhart, D.E., McClelland, J.L. in PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in Microfeatures of Cognition, vol. 1&2*. Cambridge, MA: The MIT Press.

Waterson, J. in Milmo, D. (2021). »Facebook whistleblower Frances Haugen calls for urgent external regulation«. *The Guardian* (25. okt. 2021). URL = https://www.theguardian.com/technology/2021/oct/25/facebook-whistleblower-frances-haugen-calls-for-urgent-external-regulation.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.