Izvirni znanstveni članek/Article (1.01)
Bogoslovni vestnik/Theological Quarterly 81 (2021) 4, 935—946
Besedilo prejeto/Received:09/2021; sprejeto/Accepted:09/2021
UDK/UDC: 004.89:61
DOI: 10.34291/BV2021/04/Miklavcic

Jonas Miklavčič

Zaupanje in uspešnost umetne inteligence v medicini

Trust and Success of Artificial Intelligence in Medicine

Povzetek: Evropska komisija je aprila 2021 izdala predlog pravilnika, ki bo omogočil regulacijo umetne inteligence (UI). Mnogi kompleksni algoritemski sistemi, ki v medicini ponujajo največ upanja za drastičen napredek, strogih kriterijev, ki jih v svojem predlogu navaja Evropska komisija, pogosto žal ne dosegajo – npr. ne delujejo transparentno. Kot enega izmed kriterijev za presojanje etičnosti uporabe UI tako predlagam kriterij uspešnega delovanja, saj bi morda lahko sistemom, ki konsistentno delujejo uspešno, zaupali tudi, če kriterija transparentnosti ne bi dosegali. Tu naletimo na problem odnosa med zaupanjem in uspešnostjo. Ne le, da je naše zaupanje v odnosu do UI odvisno od uspešnega delovanja sistemov, pač pa je tudi uspešnost njihovega delovanja odvisna od našega zaupanja – saj njihova odličnost temelji tudi na uporabi podatkov, ki jim jih zaupamo v učne namene. Morda je edini način za rešitev tega krožnega problema ta, da sistemom zaupamo, tudi ko ti zaupanja še niso povsem vredni.

Ključne besede: umetna inteligenca, medicina, transparentnost, zaupanje, uspešno delovanje

Abstract: The European Commission has proposed legislation to enable the regulation of artificial intelligence (AI). However, many complex algorithmic systems that offer the greatest hope for dramatic advances in medicine often fail to meet the strict criteria of the European Commission – e.g. they often do not operate transparently. As one of the criteria for assessing the ethical use of AI, I propose the criterion of successful performance since perhaps systems that work consistently well could be trusted even if they do not meet the criterion of transparency. Here we encounter the problem of the relationship between trust and performance. Not only does our trust in AI depend on the successful operation of the systems, but the success of their operation depends on our trust since their performance also relies on the use of data we provide for algorithm's learning. Perhaps we need to trust the systems before they can be trustworthy.

Keywords: artificial intelligence, medicine, transparency, trust, successful performance

1. Uvod

S povečevanjem vpliva, ki ga ima umetna inteligenca (UI) na vsakdanje življenje ljudi, se povečuje tudi potreba po njeni pravni ureditvi, ki mora – če želi biti ustrezna – temeljiti tudi na tehtnih premislekih, ki se tičejo vprašanj, vezanih na moralno filozofijo in na naravo umetne inteligence same. V prispevku želim izpostaviti specifične etične vidike, ki – če bo prikaz uspešen – očitno temeljijo na pomembnih tehničnih vidikih tehnologije, ki prodira na ključna družbena področja, tudi v medicino. Če že ne njegova vzpostavitev, pa vsaj prikaz jasne potrebe po interdisciplinarnem pristopu k reševanju situacije, v kateri smo se znašli, predstavlja glavni cilj tega prispevka.

Posebej bi rad izpostavil, da v pričujočem prispevku področje medicine služi predvsem kot platno; nanj je mogoče na relativno jasen način projicirati problemski sklop, iz katerega izhajajo vprašanja, ki jih odpiram – še zdaleč pa medicina ni edino področje, ki se ga tovrstna vprašanja tičejo.

V prispevku si bomo najprej ogledali nekaj za našo razpravo ključnih točk, ki jih Evropska komisija navaja v svojem novem predlogu pravilnika – ta naj bi predvidoma omogočil regulacijo umetne inteligence v Evropski uniji. Izpostavili bomo tudi, zakaj so nekateri v predlogu zastavljeni kriteriji, ki bi jih morali dosegati sistemi UI z visoko stopnjo tveganja, morda vendarle nekoliko prestrogi. V nadaljevanju bomo predlagali kriterij uspešnega delovanja, ki bi se morda lahko izognil kriteriju transparentnega delovanja sistemov, ki žal pogosto ni dosegljiv. V osrednjem delu prispevka bomo pokazali, kako je naše zaupanje sistemom predpogoj za njihovo uspešnost, v zadnjem pa še, da lahko to zaupanje v osnovi izhaja le iz čistega upanja, da umetna inteligenca življenja dejansko lahko rešuje. In zdi se, da moramo tu sodelovati vsi.

2. Predlog Evropske komisije

Evropska komisija je 21. aprila 2021 izdala predlog pravilnika, ki bo v prihodnje urejal področje umetne inteligence v Evropski uniji. Namen regulacije je omogočiti ‚zaupanja vredno umetno inteligenco' in zagotoviti pravni okvir za njeno etično uporabo. V prihajajočih mesecih bosta Evropski parlament in Svet predlog ločeno pregledala, dopolnila in po potrebi spremenila, nato pa sledijo še pogajanja o končnem zakonu. Dejanski ureditvi področja torej verjetno še nismo blizu, gotovo pa je nakazana smer obetavna. Storjeni prvi koraki so ključni že zato, ker odražajo jasno željo, da tako pomembno področje začnemo regulirati čim hitreje – da končno omogočimo uporabo zaupanja vrednih sistemov in hkrati onemogočimo uporabo tistih, pri katerih tveganja ne moremo upravičiti.

Predlog zakona temelji na delitvi sistemov umetne inteligence v štiri skupine,[1] ki naj bi odražale stopnjo tveganja, ki ga uporaba sistemov prinaša. Lahko si predsta-

[1] V predlogu so sistemi pravzaprav razdeljeni v tri skupine, a za lažjo preglednost sledim delitvi na štiri skupine, ki jo Evropska komisija v povzetku predloga navaja na svoji spletni strani (European Commison 2021).

vljamo piramido s štirimi nivoji. Na njenem dnu je največ sistemov – in ti po mnenju Evropske komisije predstavljajo minimalno stopnjo tveganja. V to skupino sodijo npr. umetna inteligenca v videoigrah in t. i. *spam* filtri, ki v nabiralniku naše elektronske pošte neželeno pošto ločujejo od želene (European Commision 2021). Na drugem nivoju piramide so sistemi, ki predstavljajo omejeno stopnjo tveganja in zato zanje Evropska komisija predlaga dodatno varovalo – te sisteme umetne inteligence zavezuje posebna dolžnost, ki je vezana na zagotovitev transparentnosti.[2] Tovrstni sistemi bodo morali (kljub temu, da načeloma ne predstavljajo velikega tveganja) uporabnika vsaj opozoriti na to, da ima opravka s sistemom umetne inteligence. Ko po spletu komuniciramo z banko, vse pogosteje govorimo s t. i. *chatboti* (stroji, ki simulirajo komunikacijo s človeškim svetovalcem) – in predlog zakona vključuje zahtevo, da je uporabniku eksplicitno dano vedeti, da govori s strojem in ne z živo osebo (European Commision 2021) – saj mu to omogoča, da lahko sprejme informirano odločitev, ali tak pogovor želi nadaljevati. Podobno bo veljalo tudi za t. i. ‚globoke ponaredke' (‚deepfakes') (Evropska komisija 2021, 15). Na vrhu piramide so sistemi, ki predstavljajo nesprejemljivo tveganje. Teh sistemov je sicer najmanj, a njihova uporaba varnost, dobrobit in pravice ljudi tako močno ogroža, da predvidoma ne bodo dovoljeni v nobenem kontekstu in pod nobenimi pogoji. V to skupino prepovedanih sistemov Evropska komisija uvršča npr. sisteme družbenega točkovanja[3] (Evropska komisija 2021, 22) in otroške igrače z glasovno asistenco, ki temelji na umetni inteligenci in pri otrocih lahko spodbuja nevarno vedenje (European Commision 2021).

Drugi najvišji nivo na piramidi vključuje sisteme z veliko stopnjo tveganja. Ti sistemi so za etiko umetne inteligence največji izziv. Gre za sisteme, ki so lahko odlični in posledično zelo uporabni, a prav zaradi vrste njihovega namena in področij, na katerih jih uporabljamo, lahko predstavljajo veliko stopnjo tveganja. Nekatera področja uporabe takšnih sistemov so npr. kritična infrastruktura (avtomatizirana vozila), izobraževanje (avtomatizirano ocenjevanje testov), pa tudi zaposlovanje (avtomatizirano pregledovanje in rangiranje življenjepisov) in policija (skeniranje obrazov) (European Commision 2021). V to skupino spada seveda tudi medicina (postavljanje diagnoz in odločanje o poteku zdravljenja).

Ti sistemi, ki predstavljajo visoko stopnjo tveganja, bodo za vstop v kategorijo *zaupanja vredne umetne inteligence* morali dosegati posebno stroge pogoje (Evropska komisija 2021; High-Level Expert Group on Artificial Intelligence 2019). Pogoji, ki bodo morali biti izpolnjeni, so (European Commision 2021):

– Ustrezna ocenitev stopnje tveganja in dodatni sistemi za zmanjševanje tveganja.
– Visoka kvaliteta vhodnih podatkov za minimaliziranje tveganja in diskrimina-

[2] Nekoliko nerodno je v predlogu Evropske komisije (in tudi sicer v širšem diskurzu o umetni inteligenci) uporabljen izraz ‚transparentnost', ker se uporablja za različne fenomene – kar govor o umetni inteligenci dela manj jasnega in preglednega. Kot bomo videli, na tem mestu ‚transparentnost' pomeni zagotovitev, da uporabnik ve, da ima opravka z umetno inteligenco, v nadaljevanju pa bo ‚transparentnost' zagotovitev, da uporabnik ve, kaj in kako sistem umetne inteligence dejansko deluje.

[3] Na Kitajskem v omejeni obliki že uporabljajo sistem, ki naj bi beležil, kaj prebivalci kupujejo, s kom se družijo, katere knjige berejo itd. Na podlagi ‚družbenih točk', ki so odvisne od njihovih dejavnosti, pa jim nato določene družbene ugodnosti ali pripadajo ali so jim odvzete (Kshetri 2020).

tornih rezultatov.
– Beleženje dejavnosti sistema za zagotovitev sledljivosti rezultatov.
– Podrobna dokumentacija, ki zagotavlja vse informacije o sistemu in njegovem namenu, da avtoriteta lahko oceni njuno skladnost.
– Jasne in adekvatne informacije za uporabnika.
– Ustrezen človeški nadzor za minimaliziranje tveganja.
– Visoka stopnja robustnosti, varnosti in natančnosti.

Za naš prispevek sta relevantni predvsem druga in peta točka teh sedmih zahtev. Drugi točki tu ne bomo posvečali posebne razlage, ker bo njena pomembnost ob branju prispevka postala očitna neodvisno od dodatnega pojasnila, peto točko pa si oglejmo nekoliko podrobneje. Peti kriterij, ki ga Evropska komisija predvideva kot pogoj, ki ga sistem iz skupine sistemov z veliko stopnjo tveganja mora izpolnjevati, od sistema med drugim zahteva, da uporabniku na zanj ustrezen način zagotovi transparentno poročilo o tem, kako je opravil svojo nalogo (npr. sprejel odločitev, če gre za sisteme algoritemskega odločanja). Čeprav obravnavana točka izraza ‚transparentnost' ne vsebuje, se v predlogu Evropske komisije ta izraz v povezavi s peto točko omenja pogosto.[4] V predlogu je tako zapisano:

> »Za odpravo neprepustnosti [*opacity*], zaradi katere so nekateri umetnointeligenčni sistemi fizičnim osebam morda nerazumljivi ali zanje preveč zapleteni, bi bilo treba za umetnointeligenčne sisteme velikega tveganja zahtevati določeno stopnjo preglednosti [*transparency*].« (Evropska komisija 2021, 31)

Podrobneje je peta točka predstavljena v razdelku z naslovom „Preglednost in zagotavljanje informacij uporabnikom", ki med drugim pravi:

> »Umetnointeligenčni sistemi velikega tveganja so zasnovani in razviti tako, da je njihovo delovanje dovolj pregledno [transparent], da lahko uporabniki razlagajo izhodne podatke sistema in jih ustrezno uporabijo.« (50)

Sistemi, ki zaradi področja uporabe (npr. medicina) predstavljajo veliko stopnjo tveganja, bodo tako morali delovati dovolj transparentno, da bodo uporabniki njihove izhodne podatke lahko prav interpretirali. Smiselno se zdi izpostaviti problematičnost te formulacije. Ne le, da izraz ‚dovolj pregledno' sam ne pove veliko o zahtevani stopnji preglednosti (kriterij, »da lahko uporabniki razlagajo izhodne podatke sistema«, se problemu relativnosti ne izogne), pač pa Evropska komisija predvideva tudi, da ni dovolj, da so sistemi transparentni le načeloma (npr. za programerje teh algoritemskih sistemov), ampak morajo biti dovolj transparentni, da jih lahko na pomenljiv način interpretira uporabnik sam – doseči takšno stopnjo interpretabilnosti pa je pogosto, kot bomo videli, nemogoča naloga. Še posebej nenavadno je formuliran prav problem transparentnosti. Zapisano je namreč, da mora biti *delovanje* sistema dovolj pregledno, da uporabnik lahko interpretira *rezultat*

[4] V slovenski različici predloga se uporablja izraz ‚preglednost', a sam ohranjam kar ‚transparentnost', ker je v preostali sodobni literaturi o umetni inteligenci ta izraz najbolj uveljavljen.

(izhodne podatke). Gre za dve povsem različni točki v procesu in naj poudarimo, da z interpretacijo izhodnih podatkov (rezultata) težav navadno sploh ni – težave so pri interpretaciji procesa, ki je do rezultatov vodil. Čeprav je zelo pomembna, pa lahko težavnost te formulacije v tem prispevku na srečo nekoliko zanemarimo. Ključno je le, da Evropska komisija predvideva, da je za zagotovitev zaupanja vredne umetne inteligence – kar je pogoj za njeno etično uporabo – potrebna tolikšna stopnja transparentnosti, da je sistem za uporabnika pomenljivo pregleden.

Sistemi, ki temeljijo na navadno zelo kompleksnih modelih umetne inteligence (npr. nevronske mreže), pogosto delujejo dobro, a žal ne transparentno. To preprosto pomeni, da sistemu damo določeno nalogo in učne podatke, na katerih se bo nalogo naučil opravljati (Zhau in Chen 2018) – na primer zelo veliko fotografij šimpanzov in nalogo, naj se nauči šimpanze prepoznavati. Stroj se nato sam (na) uči prepoznavati vzorce, ki mu bodo v prihodnosti omogočili na fotografijah identificirati šimpanze. V naslednjem koraku vstavimo (vhodni podatki) nabor fotografij različnih živali – in če je bilo učenje uspešno, bo iz nabora fotografij s precejšnjo natančnostjo izbral tiste, na katerih je šimpanz. Problem je, da pri nekaterih podobnih nalogah nimamo vpogleda v to, kako je prepoznavanje vzorcev med učenjem in nato izbiranje fotografij s šimpanzi dejansko potekalo. Tega vpogleda zaradi netransparentne narave kompleksnih sistemov pogosto nimajo niti sami strokovnjaki, ki so algoritem oblikovali – kaj šele uporabniki (Mittelstadt et al. 2016). V medicini prepoznavanje slik poteka na zelo podoben način kot v zgornjem primeru šimpanzov. Računalniški vid se v medicini uporablja na primer za pregledovanje rentgenskih slik, s čimer sistem, ki se je naučil prepoznavati odstopanja od normale, pomaga zdravniku pri postavitvi ustrezne diagnoze.

3. Kriterij uspešnega delovanja

Kriteriji, ki jih predlaga Evropska komisija, so tako strogi, da uporabo netransparentnih sistemov povsem izključujejo. Problem je, da so večinoma netransparentni sistemi prav tisti algoritmi, ki so sicer res najkompleksnejši, a za hiter in drastičen napredek v medicini pogosto predstavljajo največ upanja. Sistemi navadno delujejo zelo dobro – diagnoze pogosto postavljajo vsaj tako dobro kot zdravniki, praktično vedno pa hitreje (Sidney-Gibbons et al. 2019).[5] Ob upoštevanju vseh naštetih prednosti se rešitev, da jih zato, ker kriterija transparentnega delovanja ne dosegajo, preprosto ne bi uporabljali, ne zdi optimalna.

Ključno vprašanje je torej: ali obstaja kak drug kriterij, ki bi etično uporabo netransparentnih sistemov lahko omogočil? Iskanje kriterija, ki bi bil dovolj močan, da bi nekatere druge etične kriterije v njegovem imenu zanemarili, se kaže kot relativno težka naloga. Zdi se namreč, da takšnih kriterijev ni veliko, a po mojem

[5] Poznamo kar nekaj primerov, ko zdravniki sistemom niso več kos. Poleg neverjetne natančnosti in zmožnosti pregledovanja velike količine podatkov algoritemski sistemi preglede opravljajo tudi po 24 ur na dan, ne da bi (na primer med nočno izmeno) njihova uspešnost kakorkoli upadla (Gui in Chan 2017, 77).

mnenju obstaja vsaj eden – kriterij uspešnega delovanja. Kriterij uspešnega delovanja – preprosto povedano – pomeni, da če bi obravnavani sistemi konsistentno delovali uspešno,[6] reševali življenja ter celostno izboljšali kakovost življenja mnogih ljudi, morda vpogled v podrobnosti njihovega delovanja ne bi bil tako zelo ključen. Ker je uspešnost kategorija, za katero je značilen določen razpon, imamo lahko sisteme, ki so uspešni bolj ali manj. Zato bi bil verjetno dober kriterij za ‚zadostno uspešnost' ta, da algoritmi nalogo konsistentno opravljajo uspešneje kot človeški zdravniki – in to ni redkost (Gui in Chan 2017).

Predlagani kriterij uspešnega delovanja ni neproblematičen. Čeprav so sistemi umetne inteligence v medicini pogosto zelo uspešni, pa nemalokrat tudi niso – in medicina ni področje, kjer bi si lahko privoščili veliko napak. Poznamo veliko primerov, kjer so bili sistemi uporabljeni testno – in če ne bi šlo le za test, bi pacienti lahko utrpeli resne posledice, v nekaterih primerih tudi smrt.[7] Razlogov, zakaj mnogo sistemov še vedno pogosto deluje relativno neuspešno, je več. Uspešnost algoritmov je odvisna npr. od tega, ali je algoritem spisan ustrezno, da napak ne dela že kar sam po sebi. Slabo spisan algoritem bo namreč dajal slab rezultat ne glede na zelo dobro izbrane učne podatke in brezhiben način uporabe algoritma. V precejšnji meri pa je uspešnost odvisna tudi od količine in kvalitete učnih podatkov.[8] Potrebnih je namreč zelo veliko dobrih podatkov, na katerih se sistem (na)uči prepoznavati vzorce, ki jih kasneje pri svojem odločanju uporablja (Taulli 2019). Trenutno v medicini teh podatkov še ni prav veliko (Maier 2020; Willemink et al. 2020). Problem podatkov (‚the data problem') je v medicini – in tudi na mnogih drugih področjih – zaradi obsega problemov celo svoje področje raziskovanja. Jennifer Bresnick med problemi navaja nekaj največjih: podatke v medicini je težko pridobiti, izčistiti, shraniti, zavarovati, nato za njih skrbeti, jih posodabljati in uporabljati (Bresnick 2017) – problematično je lahko marsikaj. Ustrezno shranjevanje, varovanje in posodabljanje so kot problemi sicer pomembni, a ker so tehnične narave, bodo glede na hitrost tehnološkega napredka, ki smo mu priča v sodobni družbi, po mojem mnenju v bližnji prihodnosti razrešeni. Pridobivanje podatkov je medtem bolj zapleteno vprašanje. Ne le zato, ker so od rešitve tega problema odvisni vsi nadaljnji koraki in z njimi povezana problematika, temveč ker ima (ob zanimivi tehnični plati)[9] problem pridobivanja podatkov tudi pomembno etično komponento. Eden od razlogov, da je do dobrih podatkov v medicini težko priti, je namreč tudi, da so učni podatki v medicini osebne

---

[6] Kot ‚uspešno delovanje' razumem predvsem delovanje sistema, ki naloge opravlja v skladu z zastavljenim ciljem. Če sistem, katerega naloga je, da na fotografijah živali išče fotografije šimpanzov, te fotografije tudi dejansko poišče, je nalogo opravil uspešno – in upravičeno trdimo, da deluje uspešno. Izbrani izraz namenoma uporabljam zato, ker ga lahko navežemo na t. i. ‚success rate', ki je v tu nakazanem smislu objektivno izmerljiv.

[7] IBM-ov *Watson Health's Algorithm* za prognozo pri rakavih obolenjih je pri testnih poskusih zelo pogosto predlagal načine zdravljenja, ki so se izkazali za hudo napačne in bi morda vodili v smrt pacienta (Grote in Bernes 2020, 208).

[8] »Garbage in, garbage out problem« govori o tem, da slabi učni podatki vodijo v slabe rezultate ne glede na brezhibnost algoritma.

[9] Pridobivanje podatkov je tehnično zahtevno med drugim zato, ker podatki pogosto niso dostopni (lahko jih preprosto ni, lahko pa obstajajo, a ne v digitalni obliki) ali pa niso strukturirani na način, primeren za algoritmično obdelavo (Willemink 2020).

narave (saj gre za zdravstvene kartoteke) in tako zavarovani s pravicami pacientov (Maier 2020). Ljudje osebnih podatkov v učne namene sistemov prostovoljno ne želijo (ali pa potencialno ne bi želeli) zaupati. V nedavni raziskavi, ki jo je izvedlo podjetje KPMG, se je izkazalo, da približno 91 % anketirancev s področja zdravstva verjame, da bo umetna inteligenca področje medicine sicer res drastično izboljšala, a kar 75 % vprašanih je prepričanih, da bo morda istočasno ogrozila zasebnost pacientov (McGrail 2020). Nenaklonjenost prostemu deljenju svojih osebnih podatkov je seveda pričakovana in razumljiva. Pomislimo, da poleg rabe izraza ‚zaupanje' v smislu, da zaupamo *nekomu* ali *nečemu*, poznamo v slovenščini tudi rabo, da nekomu zaupamo *nekaj*. Osebi A zaupam ključe svojega avtomobila, če osebi A zaupam – če se mi kaže kot zaupanja vredna. In povsem razumljivo je, da ljudje svojih osebnih podatkov sistemom, ki trenutno – vsaj po kriterijih Evropske komisije – še niso zaupanja vredni (ker na primer ne delujejo transparentno), ne želijo zaupati.

Preden nadaljujemo s prikazom nekaterih posledic tega, da ljudje svojih podatkov v učne namene sistema ne želijo zaupati, naj omenimo še, da bi pacienti prav mogoče to bili brez težav pripravljeni storiti, če bi iz nabora podatkov, ki jih zdravstvena kartoteka vključuje, izbrisali njihovo ime (Maier 2020). Če iz zdravstvene kartoteke ni razvidno, da gre za moje podatke, ti podatki v pomembnem smislu dejansko niso več moji – in mnogim bi bilo verjetno precej vseeno, kaj se s takšnimi podatki dogaja. Zdi se, da je tako težava pravzaprav rešljiva dokaj preprosto, a posebej moramo poudariti dva pomembna vidika oz. zadržka. Prvi je tehnične narave. Sodobni algoritemski sistemi so dovolj napredni, da v večini podobnih primerov, ko jim manjka en podatek (v tem primeru ime lastnika kartoteke), na voljo pa je dovolj drugih, manjkajoči podatek brez težav poiščejo (Taulli 2019, 35). Brž ko ima stroj dostop do neke širše baze podatkov, zakritje imena brez zakrivanja tudi npr. datuma in kraja rojstva ne bo zadostno, da stroj kartoteke ne bi mogel povezati z mojim imenom. Študija, opravljena na MIT-ju, je pokazala, da so podatki, ki se zdijo relativno anonimni (ali pa jih anonimne skušamo narediti), pogosto vse prej kot to. Raziskovalci so ugotovili, da je rekonstruirati manjkajoče podatke – in identificirati posameznike – relativno lahko. Pri tem so uporabljali sistem združevanja dveh različnih naborov podatkov. Na primeru Singapurja so tako uporabili podatke iz mobilne mreže (predvsem GPS-sledenja mobilnim napravam) in podatke lokalnega transportnega sistema – po enajstih tednih analiz je raziskovalcem uspelo identificirati kar 95 % posameznikov, ki so uporabljali Singapurski transportni sistem (35). Izbris imena iz zdravstvene kartoteke tako po vsej verjetnosti – če je naš cilj, da kartoteka gotovo ostane za vedno anonimna –ne doseže veliko. Drug pomemben vidik pa ni tehnične narave, temveč se dotika dejstva, da to, da bi ob zakritju imena lastnika kartoteke ljudje svoje podatke strojem bili pripravljeni zaupati, lahko razumemo, da bi zaupanje strojem bilo možno, če določeni vhodni podatki ne bi bili transparentni. Zanimiv obrat se torej skriva v tem, da bi ravno določena stopnja netransparentnosti morda lahko omogočila celo večjo stopnjo zaupanja. Podobnih primerov, kjer popolna transparentnost povzroča nezaupanje, je veliko – pomislimo lahko na spletno plačevanje in podobne primere, kjer je skritost določenih podatkov prav pogoj za naše zaupanje. To izpostavljam zato, ker je po mojem mnenju na mestu posebej poudariti, da

sta transparentnost in zaupanje vsaj včasih v obratnosorazmernem odnosu – in cilj, da moramo zagotoviti kar največjo stopnjo transparentnosti, če želimo zaupanja vredno umetno inteligenco, vsaj ni absolutna zapoved, vsaj ne drži vselej, če ni celo kar vprašljiv. Zdi se, da ne zgrešimo preveč, če rečemo celo, da vsaj v nekaterih primerih, ko je npr. govora o osebnih podatkih in o ogrožanju zasebnosti, popolna transparentnost lahko vodi v neetično uporabo umetne inteligence.

4. Uspešnost in zaupanje

Ljudje torej svojih osebnih podatkov ne želijo (ali pa v prihodnosti, ko bo to množično potrebno, morda ne bodo želeli) ponuditi v učne namene strojev (algoritemskih sistemov). Pri tehtanju ustreznosti predlaganega kriterija uspešnega delovanja sistemov in dejstva, da ljudje sistemom UI še ne zaupajo povsem, se hitro izkaže, da sta uspešnost in zaupanje v nekaterih primerih v zelo zanimivem krožnem odnosu. Naj ponazoritev tega odnosa začnem drugje. Zelo velik problem, s katerim se soočajo v ZDA, je razreševanje umorov. Od šestdesetih let prejšnjega stoletja delež umorov, ki so bili uspešno raziskani (storilca so odkrili, prijeli in predali sodstvu), vztrajno pada iz leta v leto (Hargrove 2019).[10] Statistika je precej nenavadna glede na to, da se zdi, da smo vse bolj v dobi, ki zaradi našega puščanja digitalnih sledi na vsakem koraku dobro zakritje naših dejavnosti vedno bolj onemogoča. Razloge za upadanje uspešno raziskanih umorov je tako mogoče iskati v mnogočem. Vsaj delno lahko morda predvidevamo, da je bilo v šestdesetih letih prejšnjega stoletja več primerov, ko je bil nekdo (lahko) obtožen po krivem – po nesreči ali namenoma (pomislimo na primere, ki so podobni t. i. ‚šerifovemu scenariju' H. J. McCloskeya, ki ga poznamo v moralni filozofiji). Morda je bilo veliko umorov izpeljanih tudi manj sofisticirano. A mnogi trdijo tudi, da je vsaj delno problem v tem, da ljudje policiji ne zaupajo več zaradi njene neuspešnosti (Bier 2018; Sweeney in Groner 2020). V Chicagu, ki povprečje neraziskanih umorov v ZDA drastično znižuje, kar polovica ljudi, ki so bili ustreljeni in so preživeli, po dogodku o njem s policijo ne želi govoriti (Bier 2018). Ker ne morejo pričakovati, da bo policija storilca uspešno prijela, s svojim pogovorom s policijo (in tako kršenjem osnovnih ‚uličnih pravil') žrtve enostavno tvegajo preveč. Zanimivo je, da nekateri avtorji navajajo, da je chicaška policija med drugim tako zelo neuspešna vsaj delno tudi zato, ker ji ljudje pač ne zaupajo (Bier 2018). Gre za relativno težko rešljiv problem – saj ljudje policiji ne zaupajo zaradi njene neuspešnosti, neuspešnost policije pa je vsaj delno tudi posledica nezaupanja ljudi.

Pri algoritmih v medicini naletimo na zelo podobno težavo, le da sta tu uspešnost in zaupanje prepletena še bolj neposredno in tesneje. Sistemom – povsem upravičeno – še ne zaupamo zaradi njihove pogoste neuspešnosti (ali vsaj vprašljive uspešnosti), vsaj delno pa je njihova neuspešnost tudi posledica ravno na-

[10] Statistika od vira do vira nekoliko variira (tudi, ker pogosto ni poenoteno, katere vrste ubojev/umorov so v statistiko vključene, in posebej še, kakšni so kriteriji, da umor velja za ‚raziskan'), a številke večinoma ne odstopajo veliko. Leta 1965 je bilo v ZDA raziskanih približno 83 % umorov, leta 1975 78 %, leta 1985 72 %, 1995 65 %, 2005 62 %, 2019 58 % (Hargrove 2019).

šega nezaupanja: ker je uspešnost delovanja sistema odvisna od količine in kvalitete učnih podatkov, ki jih v učne namene algoritmov zaupamo, s svojim nezaupanjem uspešnost njihovega delovanja neposredno zaviramo. Povedano preprosteje: ne zaupamo jim, ker niso dovolj uspešni, in vsaj delno niso uspešni, ker jim še ne zaupamo dovolj. Kako torej rešiti ta krožni problem? Zdi se, da ga bomo pri algoritmih rešili le stežka, saj ti brez naših podatkov, ki bi jih uporabili v učne namene, enostavno ne bodo nikoli dovolj uspešni, da bi jim lahko upravičeno zaupali. Ne ostaja nam drugega, kot da krog prekinemo sami – pri nezaupanju.

V skladu s povedanim trdim, da če želimo povsem zaupanja vredno umetno inteligenco, ji bomo morda morali zaupati tudi, ko zaupanja še ne bo povsem vredna.

5. Upanje kot ključni vir zaupanja

Reševanje problema, kako zaupati stvari, ki zaupanja še ni vredna, je seveda težavna naloga. Eden od možnih pristopov bi lahko bil, da začnemo zaupanje do umetne inteligence graditi postopoma, po korakih. V prvem koraku bi lahko poskusili pridobiti splošno zaupanje do umetne inteligence pri sistemih, ki so vsaj načelno pregledni, transparentni – da bi ljudje videli, kako zelo so nam lahko v pomoč in da – vsaj v osnovi – hujših groženj ne predstavljajo. V tem smislu se nekoliko strožji kriteriji, kot jih predlaga Evropska komisija, dejansko kažejo kot zelo uporabni. V naslednjem koraku lahko sledi postopno testno uporabljanje sistemov v medicini. V tem koraku ti sistemi o ničemer še ne odločajo, le vzporedno in neodvisno od zdravnikovega odločanja jih uporabljamo toliko, da preverjamo njihovo uspešnost. Morda v tretjem koraku nato sledi uporaba sistemov na področjih medicine, kjer posledice napak niso velike in si jih nekaj morda lahko privoščimo. In tako naprej.

Seveda je problem kompleksnejši, ker tudi drugi dejavniki, npr. dodelitev odgovornosti, ko gre pri odločanju nekaj hudo narobe,[11] in druge pravne ureditve, za katere moramo še poskrbeti, morda kažejo na dodatne korake k popolnemu zaupanju. Vojko Strahovnik poleg potrebnosti ureditve pravnega okvira odgovornosti v primeru napak dobro izpostavlja še, da transparentnost v primeru zelo dobro delujočih algoritmov morda res ni ključna, a gotovo ni zanemarljiva – in vseeno ostaja ideal, kolikor je pač dosegljiva:

> »/… / vsekakor pa se je smiselno vprašati, kakšno raven transparentnosti lahko v teh postopkih od takšnih sistemov pričakujemo (kljub njihovi zanesljivosti oziroma natančnosti, ki morda prekaša človeške odločevalce) in kako je z odgovornostjo v primeru napak oziroma zmot.« (Strahovnik 2019, 602)

V vsakem primeru bo po mojem mnenju – ne glede na postopnost grajenja zaupanja – na določeni točki potreben nek *leap of faith*. Neizogibno se mi torej zdi,

[11] V predlogu Evropske komisije je zaenkrat predvideno, da splošno odgovornost za dajanje sistemov na trg in v uporabo prevzame ponudnik sistema, tudi če sistema ponudnik ni zasnoval ali razvil sam (Evropska komisija 2021, 32).

da se nam bo umetna inteligenca, ker je pogosto odvisna od našega zaupanja, morala vsaj zdeti zaupanja vredna, *preden* in *da sploh* bo lahko zaupanja vredna. In kot kaže, se bo to naše uvodno, še neupravičeno zaupanje moralo napajati iz čistega upanja, da umetna inteligenca življenja dejansko lahko rešuje – ali vsaj izboljša njihovo kakovost.

6. Neizogibnost netransparentnosti?

Kljub temu, da primerov, ko nekomu ali nečemu moramo zaupati, še preden se ima ta priložnost izkazati za zaupanja vrednega, ne primanjkuje,[12] pa kaže dodatno izpostaviti posebnost primerov umorov v Chicagu in algoritmov v medicini. Ne gre torej za to, da bi na vprašljivo zaupanje pristali vnaprej in povratno informacijo, ali je oseba/sistem zaupanja vreden (in je bilo torej naše zaupanje upravičeno), dobili kasneje, pač pa za to, da z vnaprejšnjim vložkom zaupanja, ki *zagotovo še ni upravičeno*, prav *omogočimo*, da bo v prihodnosti naše zaupanje lahko upravičeno. Pri algoritmih prav to, da jim zaupamo, ko jim še ne bi smeli, omogoča, da lahko postanejo zaupanja vredni.

Če koncept transparentnosti razumemo nekoliko širše, kot le dejstvo, da jasno in razločno vidimo, *kaj dejansko je na delu*, se pojavi zanimiv problem. Ker pri algoritmih nekoliko lažna percepcija situacije (najprej se pretvarjamo, kot da jim lahko zaupamo) omogoča, da ta ista percepcija začne odražati realnost in ni več lažna (končno lahko priznamo, *kaj dejansko je na delu*), vidimo, da če je vsaj včasih za zagotovitev upravičenega zaupanja najprej nujna določena mera neupravičenega zaupanja, je prav za zagotovitev transparentnosti najprej nujna določena mera netransparentnosti. To pa nas vrača na že omenjeni odnos med transparentnostjo algoritmov in njihovo etično uporabo. Transparentno delovanje algoritmov pomeni, da imamo vpogled v to, kaj se s podatki, ki jih pri svojem odločanju algoritem uporablja, med procesom odločanja dejansko dogaja. V praksi to po navadi pomeni, da vidimo, katerim podatkom algoritem med svojim odločanjem pripisuje težo in kolikšna ta teža je. To pa namiguje, da morajo biti podatki vsaj načelno vidni. Kot sem nakazal že zgoraj, je za zaupanje včasih torej morda res potrebno, da niso vidni prav vsi podatki – in to je dober znak, da zgornja izpeljava, da je morda določena stopnja netransparentnosti (če v prihodnje želimo povsem zaupanja vredno in transparentno UI) nujna postojanka na poti do izgradnje našega upravičenega zaupanja do teh naprednih sistemov, ni daleč od resnice.

7. Zaključek

Zdravstvene krize, kakršna je tudi pandemija COVID-19, nas vse hitreje silijo v globok premislek o vlogi, ki bi jo pri reševanju življenj lahko imela najnovejša tehno-

[12] Eric M. Uslaner takšnemu zaupanju pravi ‚strateško zaupanje', saj je odločitev, da nekomu bomo zaupali, v tem smislu v osnovi strateška (Uslaner 2002, 17).

logija. Zastavljajo se tudi bolj temeljna vprašanja, ki se tičejo denimo odnosa med novo tehnologijo in telesom (Štivić 2020) ter vprašanja, ali in kako se lahko kriterij transparentnosti v kriznih situacijah spreminja (Strahovnik et al. 2020). Ravno v kriznih situacijah, ko so v igri naša življenja, se namreč kakršna koli možnost drastičnega napredka kaže kot smer, ki ji moramo slediti.

A zasledovanje napredka seveda ne sme potekati za vsako ceno in na kakršen koli način. Umetna inteligenca res neizogibno prihaja na ključna področja človeškega delovanja, vključno z medicino, in Michael Frosting se ne moti, ko pravi:

> »Dejstvo je, da – tako kot tisti, ki so izdelovali konjske vprege, niso bili vprašani, ali jim je bil všeč pojav avtomobilov – tudi nas kot zdravnikov, ki pregledujemo slike, ne bodo vprašali, ali nam je ta sprememba všeč ali ne. Odločimo se lahko le, ali želimo pri spremembi sodelovati ali pa jo poskusimo ignorirati.« (Forsting 2017, 358)

Tudi humanistika se lahko odloči le, ali želi pri spremembi sodelovati. Verjetno ne moremo vplivati na to, ali se sprememba bo zgodila, lahko pa s tehtnim premislekom prispevamo k zarisovanju čim bolj ustreznih smeri in načinov implementacije kompleksnih sistemov UI. Nenazadnje to sovpada tudi s ciljem teologije, kakor ga razume Robert Petkovšek: »Cilj, ki si ga zastavlja, ni obvladovati sodobno kulturo, ampak spodbujati univerzalne človeške vrednote in delo za skupno dobro.« (2019, 29) Cilj je torej skupno dobro in naš prispevek zagovarja, da je premišljeno sodelovanje pri etični implementaciji sistemov umetne inteligence v medicino pot do hitrega napredka in s tem do skupnega dobrega. Roman Globokar se strinja:

> »Ponovno poudarjamo, da ne nasprotujemo novim tehnologijam /…/, če le te tehnologije podpirajo celostni razvoj posameznika in družbe in ne ogrožajo sedanjega in prihodnjega pristnega človeškega življenja, ki se izraža v avtonomiji, svobodi, osebni odgovornosti in v zmožnosti za sodelovanje pri gradnji skupnega dobrega v družbi.« (2019, 627)

Vstopa umetne inteligence v medicino torej ne moremo preprečiti in menim, da ga tudi ne bi smeli želeti preprečiti – lahko pa filozofi in teologi svojo dolžnost prepoznavamo v tem, da poskusimo sodelovati pri izbiri ustreznosti njenega vstopa in njenih etičnih omejitev.

Če bo naš prispevek vsaj delno uspešen, pa se bo izkazalo tudi, da bodo morali pri implementaciji umetne inteligence na ključna področja družbenega življenja na koncu sodelovati vsi. Brezbrižnost in morda celo upor nista (več) na mestu. Na vrsti je dialog.

Kratica

UI – Umetna inteligenca.

Reference

Bier, Daniel. 2018. Why are unsolved murders on the rise? Freethink. https://www.freethink.com/social-change/why-don-t-we-solve-murder-anymore (pridobljeno 26. julija 2021).

Bresnick, Jennifer. 2017. Top 10 Challenges of Big Data Analytics in Healthcare. Health ITanalytics, 12. 6. https://healthitanalytics.com/news/top-10-challenges-of-big-data-analytics-in-healthcare (pridobljeno 28. julija 2021).

European Commission. 2021. Shaping Europe's digital future: Regulatory framework proposal on Artificial Intelligence. Directorate-General for Communications Network. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (pridobljeno 30. julija 2021).

Evropska komisija. 2021. Uredba Evropskega parlamenta in Sveta o določitvi harmoniziranih pravil o umetni inteligenci (akt o umetni inteligenci) in spremembi nekaterih zakonodajnih aktov unije. EUR-Lex, 21. 4. https://eur-lex.europa.eu/legal-content/SL/TXT/HTML/?uri=CELEX:52021PC0206&from=SL (pridobljeno 30. 6. 2021).

Forsting, Michael. 2017. Machine Learning Will Change Medicine. *JNM – The Journal of Nuclear Medicine* 58:357–358.

Globokar, Roman. 2019. Normativnost človeške narave v času biotehnološkega izpopolnjevanja človeka. *Bogoslovni vestnik* 79, št. 3:611–628.

Grote, Thomas, in Philipp Berens. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics.* 46:205–211.

Gui, Chloe, in Victoria Chan. 2017. Machine learning in medicine. *Medicine & Technology.* 86:76–78.

Hargrove, Thomas. 2019. Uniform Crime Report for Homicides: 1965–2019. Project: Cold Case. https://projectcoldcase.org/cold-case-homicide-stats/ (pridobljeno 26. julija 2021).

High-Level Expert Group on Artificial Intelligence. 2019. Shaping Europe's digital future: Ethics Guidelines for Trustworthy. Directorate-General for Communications Network, 8. 4. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Pridobljeno 8. 4. 2021).

Kshetri, Nir. 2020. China's Social Credit System: Data, Algorithms and Implications. *IT Professional* 22, št. 2:14–18.

Maier, Andreas. 2020. Will we ever solve the Shortage of Data in Medical Applications? Towards data science. https://towardsdatascience.com/will-we-ever-solve-the-shortage-of-data-in-medical-applications-70da163e2c2d (pridobljeno 28. 7. 2021).

McGrail, Samantha. 2020. Virtualization News: Challenges of Artificial Intelligence Adoption in Healthcare. HITInfrastructure, 14. 2. https://hitinfrastructure.com/news/challenges-of-artificial-intelligence-adoption-in-healthcare (pridobljeno 28. 7. 2021).

Mittelstadt, Daniel Brent, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter in Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, št. 2:1–21.

Petkovšek, Robert. 2019. Teologija pred izzivi sodobne antropološke krize: preambula apostolske konstitucije Veritatis gaudium. *Bogoslovni vestnik* 79, št. 1:17–31.

Sidney-Gibbons, Jenni A. M. in Chris J. Sidney-Gibbons. 2019. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology* 19:64.

Strahovnik, Vojko. 2019. Vrline in transhumanistična nadgradnja človeka. *Bogoslovni vestnik* 79, št. 3:601–610.

Strahovnik, Vojko, Jonas Miklavčič in Mateja Centa. 2020. Etični vidiki uporabe algoritemskega odločanja in ostalih sistemov UI v času pandemij oz. izrednih razmer. *Bogoslovni vestnik* 80, št. 2:321–334.

Sweeney, Annie, in Jeremy Groner. 2020. Chicago police's homicide clearance rate dips in 2020 after improvement in recent years. Chicago Tribune, 14. 12. https://www.chicagotribune.com/news/criminal-justice/ct-chicago-police-2020-clearance-rates-20201215-2evyuaybxbcvxex7s4wlvrx62q-story.html (pridobljeno 26. 7. 2021).

Štivić, Stjepan. 2020. Body in Temptation: An Attempt at Orientation in a Boundary Situation. *Bogoslovni vestnik* 80, št. 2:443–451.

Taulli, Tom. 2019. *Artificial Intelligence Basics: A Non-Technical Introduction.* New York City: Apress.

Uslaner, Eric M. 2002. *The Moral Foundations of Trust.* Cambridge: Cambridge University Press.

Willemink, Martin J., Wojciech A. Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, Daniel L. Rubin in Matthew P. Lungren. 2020. Preparing Medical Imaging Data for Machine Learning. *Radiology* 295, št. 1:4–15.

Zhou, Jianlong, in Fang Chen, ur. 2018. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent.* Berlin: Springer.