Osnaživanje povjerenja osiguravanjem pravednosti u sustavima umjetne inteligencije u zdravstvenoj skrbi

Luka Poslon*
luka.poslon@unicath.hr
https://orcid.org/0000-0002-7389-7694

https://doi.org/10.31192/np.23.2.4
UDK: 614:004.8
004.8:614]:177.7
Pregledni članak / Review
Primljeno: 27. veljače 2025.
Prihvaćeno: 5. svibnja 2025.

Rad se bavi suvremenim izazovima pravednosti u kontekstu primjene umjetne inteligencije (UI) u zdravstvenoj skrbi. Osim izazova konceptualizacije pravednosti, prikazana je i važnost objašnjivosti i objašnjive umjetne inteligencije putem koje se osigurava povjerenje u primjenu UI u zdravstvu. Neželjeni primjeri primjene UI u zdravstvu podsjećaju da je moguće da sustav UI predlože nepravedne i pristrane ishode predviđanja koja mogu dodatno produbiti društvene nejednakosti na rasnoj, etničkoj ili drugim pripadnostima. Potrebno je stoga razvijati sustave UI koji će ublažavati pristranost i posljedice diskriminacije koji mogu nepovoljno utjecati na procese donošenja odluka. Usprkos alatima poput »AI Fairness 360« i »What-If Tool« koji pružaju tehnička rješenja za smanjenje pristranosti u sustavima UI, potrebno je uložiti dodatan napor u razvoj UI koja pruža mogućnost objašnjenja uz ispunjenje etičkih normi. Stoga se rješenje za ublažavanje nepravednosti i pristranosti može naći u sustavu UI pod nazivom TWIX koji, zahvaljujući mogućnostima objašnjivosti, osigurava pravednija predviđanja. Buduća istraživanja trebala bi biti usmjerena u razvoj sličnih sustava UI s mogućnostima objašnjenja predviđanja radi pravednijega donošenja odluka.

Ključne riječi: *objašnjivost, povjerenje, pravednost, umjetna inteligencija, zdravstvena skrb.*

* Luka Poslon, mag. phil., Hrvatsko katoličko sveučilište, Laboratorij za etiku digitalnih tehnologija u zdravstvu (Digit-HeaL), Ilica 244, HR-10000, Zagreb.

Uvod

Pravednost je jedan od najvažnijih principa kojim se stručnjaci vode prilikom donošenja odluka za dijagnozu, prognozu ili terapiju bolesti te ima vodeću ulogu u kontekstu zdravstvene skrbi.[1] Sustavi umjetne inteligencije (dalje: UI) se danas koriste u brojnim osjetljivim okruženjima pri donošenju važnih odluka, na primjer u javnoj upravi,[2] pravnom sustavu[3] ili zapošljavanju.[4] Posljednjih godina sve se više naglašava unapređenje poštivanja etičkih i epistemoloških načela u radu sustava UI, što je uzrokovalo da sve više znanstvenika radi na razvoju sustava UI u skladu s tim načelima.[5]

Zbog sve veće primjene UI u zdravstvenoj skrbi, koja pokušava poboljšati kvalitetu i učinkovitost zdravstvene skrbi te odgovoriti na nedostatak liječnika i sve veći broj pacijenata, UI je sve više uključena u procese donošenja odluka. To može dovesti do bržih i boljih ishoda odlučivanja.[6] Poznati su brojni primjeri primjena sustava UI koji pomažu prilikom regulacije pravovremene intervencije i prevencije nepovoljnih ishoda u jedinicama intenzivnog liječenja, što je teško postići bez njih s obzirom na ograničene ljudske sposobnosti istovremenog obavljanja više zadataka uz optimalnu učinkovitost. Takvi primjeri uključuju upotrebu algoritama i statističkih tehnika za poboljšanje sposobnosti računala da predlože bolje odluke temeljene na sposobnosti brze analize, razumijevanja i interpretacije velike količine podataka.[7]

Kada se sustavi UI koriste u kontekstu liječenja pacijenata bolnice, zdravstvene ustanove i istraživači moraju biti sigurni da neće nastati neočekivane društvene implikacije, poput pristranosti prema rasi, etničkoj pripadnosti i/ili

---

[1] Usp. Alessa ANGERSCHMID i dr., Fairness and Explanation in AI-Informed Decision Making, *Machine Learning and Knowledge Extraction*, 4 (2022) 556-579; María Agustina RICCI LARA i dr., Addressing fairness in artificial intelligence for medical imaging, *Nature Communications*, 13 (2022) 1-6; Min Kyung LEE, Kate RICH, Who Is Included in Human Perceptions of AI?. Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (2021) 1-14.

[2] Usp. Ignacio CRIADO, Rodrigo SANDOVAL-ALMAZÁN, J. Ramon GIL-GARCIA, Artificial intelligence and public administration. Understanding actors, governance, and policy from micro, meso, and macro perspectives, *Public Policy and Administration*, (2024) 1-12, Special Issue: Artificial intelligence and public administration: actors, governance, and policies https://doi.org/10.1177/09520767241272921.

[3] Usp. Alexandra CHOULDECHOVA, Fair Prediction with Disparate Impact. A Study of Bias in Recidivism Prediction Instruments, *Big data*, 5 (2017) 2, 153-163, doi:10.1089/big.2016.0047.

[4] Usp. Yalcin ACIKGOZ i dr., Justice Perceptions of Artificial Intelligence in Selection, *International Journal of Selection and Assessment*, 28 (2020) 4, 399-416.

[5] Usp. Daiju UEDA i dr., Fairness of artificial intelligence in healthcare: review and recommendations, *Japanese Journal of Radiology*, 42 (2024) 3-15, https://doi.org/10.1007/s11604-023-01474-3.

[6] Usp. Bruno LEPRI i dr., Fair, Transparent, and Accountable Algorithmic Decision-making Processes, *Philosophy & Technology*, 31 (2017) 611-627.

[7] Usp. Vinay SURESH i dr., Artificial Intelligence in the Intensive Care Unit. Current Evidence on an Inevitable Future Tool, *Cureus*, 16 (2024) 5, 2-11, doi:10.7759/cureus.59797.

osobama s invaliditetom. Nažalost, u razvoj algoritama UI uključeni su razni oblici pristranosti koji mogu značajno utjecati na pravednost pogrešnim prikazivanjima rezultata predviđanja.[8]

U području zdravstvene skrbi uzroci pristranosti i diskriminacije u rezultatima predviđanja UI su već istraživani s naglaskom na osiguravanje distributivne pravednosti koja osigurava jednakost u učinku i raspodjeli resursa[9] te s naglaskom na posljedice nepravednih predviđanja koja utječu na dijagnostičke netočnosti za različite društvene skupine.[10]

Usprkos široko raširenoj primjeni u zdravstvu, pojedini sustavi UI rade na principu »crne kutije« i algoritamske neprozirnosti, što znači da ne razumijemo mehanizme rada niti kako točno UI donosi odluke.[11] Razumijevanje mehanizama rada sustava UI koji izrađuju predviđanja vrlo je važno u procjeni povjerenja, posebice kada liječnici ili zdravstveno osoblje svoju odluku za daljnje medicinske postupke temelje na takvom predviđanju sustava UI ili u prilikama kada odlučuju hoće li za odluku konzultirati predviđanja sustava UI. Takvo razumijevanje pravednosti naglašava nužnost uvida u mehanizme rada sustava UI, koji se metodama objašnjive umjetne inteligencije (dalje: OUI) mogu unaprijediti putem uključivanja objašnjenja vlastitih predviđanja u sustave UI koji se tada koriste s povjerenjem za razliku od onih koji ne nude mogućnost vlastitih objašnjenja.

No, s obzirom na sve veću složenost metoda sustava UI putem kojih nastaju predviđanja, potrebno je razvijati i načine objašnjenja tih predviđanja da bi se moglo nastaviti razvijati povjerenje u rad sustava UI. Zbog toga se OUI smatra potencijalnim rješenjem za ublažavanje algoritamske neprozirnosti, na temelju ključne uloge u poboljšanju objašnjenja donošenja odluka podržanih UI-om.[12] S obzirom na povezanost objašnjivosti i povjerenja, neki znanstvenici tvrde da je nužno izgrađivati OUI da bi se moglo izgrađivati povjerenje u primjenu UI.[13]

Cilj rada je prikazati izazove konceptualizacije pravednosti u kontekstu zdravstvene skrbi te ponuditi zadovoljavajuću definiciju pravednosti. S obzirom na primjenu sustava UI u zdravstvu, u ovom se radu ističe mogućnost nepravednih i pristranih ishoda predviđanja koji mogu nepovoljno utjecati na

[8] Usp. Kristine BÆRØE i dr., Can medical algorithms be fair? Three ethical quandaries and one dilemma, *BMJ health & care informatics*, 29 (2022) 1, 1-6, e100445, doi:10.1136/bmjhci-2021-100445.

[9] Usp. Alvin RAJKOMAR i dr., Ensuring Fairness in Machine Learning to Advance Health Equity, *Annals of internal medicine*, 169 (2018) 12, 866-872, doi:10.7326/M18-1990.

[10] Usp. Qizhang FENG i dr., Fair Machine Learning in Healthcare. A Survey, *IEEE Transactions on Artificial Intelligence*, (2024) 1-16.

[11] Usp. Pantelis LINARDATOS, Vasilis PAPASTEFANOPOULOS, Sotiris KOTSIANTIS, Explainable AI: A Review of Machine Learning Interpretability Methods, *Entropy*, 23 (2020) 18, 1-45, doi:10.3390/e23010018.

[12] Usp. David GUNNING i dr., DARPA's explainable AI (XAI) program. A retrospective, *Applied AI Letters*, 2 (2021) 1-11.

[13] Usp. Jan BEGER, The crucial role of explainability in healthcare AI, *European Journal of Radiology*, 176 (2024) 111507, 1-2, doi:10.1016/j.ejrad.2024.111507.

procese donošenja odluka. Predstavit će se moguća rješenja za ublažavanje nepravednosti i pristranosti u sustavima UI. Konačno, prikazat će sustav UI TWIX[14] koji objašnjivošću osigurava pravednija predviđanja i ublažava pristranosti zbog čega bi se u budućnosti trebalo nastaviti s razvojem sličnih sustava UI u zdravstvenoj skrbi zbog mogućnosti pravednijeg donošenja odluka.

1. Izazovi konceptualizacije pravednosti u kontekstu zdravstvene skrbi

Pravednost se u literaturi ističe kao jedno od temeljnih načela etike UI,[15] a posljednjih nekoliko godina se dodatno istražuje pravednost u kontekstu UI te sve više autora naglašava potrebu za unapređenjem pravednosti[16] i popratnih fenomena kao što su objašnjivost,[17] povjerenje,[18] transparentnost[19] ili donošenje odluka pomoću sustava UI.[20] Izrađeni su i brojni pristupi prema intervencijama koje se temelje na definicijama i metrikama za matematički prikaz pristranosti, pravednosti i diskriminacije za razna područja, kao što su računalne znanosti[21] ili kazneno pravo,[22] ali im nedostaje dosljednost i sustavan pristup u konvencijama imenovanja.[23] Vrijedi napomenuti i da danas postoji mnogo različitih tumačenja pravednosti, no ne postoji univerzalan alat za njezino mjerenje.[24] Osim toga, važno je napomenuti da kvantitativne metode nisu jedini pristup

[14] Usp. Dani KIYASSEH i dr., A vision transformer for decoding surgeon activity from surgical videos, *Nature biomedical engineering*, 7 (2023) 6, 780-796, doi:10.1038/s41551-023-01010-8.

[15] Usp. Brent Daniel MITTELSTADT i dr., The Ethics of Algorithms. Mapping the Debate, *Big Data & Society*, 3 (2016) 2, 1-21; Andreas TSAMADOS i dr., The ethics of algorithms. Key problems and solutions, *AI & SOCIETY*, 37 (2020) 1-35.

[16] Usp. Reuben BINNS, Fairness in Machine Learning: Lessons from Political Philosophy, *Decision-Making in Computational Design & Technology eJournal*, 81 (2018) 149-159.

[17] Usp. Jakob SCHOEFFER, Maria DE-ARTEAGA, Niklas KÜHL, Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making, u: Florian Floyd Mueller i dr. (ur.), *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, Association for Computing Machinery, 2022, 1-18.

[18] Usp. Juan Manuel DURÁN, Nico FORMANEK, Grounds for Trust. Essential Epistemic Opacity and Computational Reliabilism, *Minds and Machines*, 28 (2018) 645-666.

[19] Usp. Andrew SELBST i dr., Fairness and Abstraction in Sociotechnical Systems, u: Jamie H. Morgenstern (ur.), *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, Association for Computing Machinery, 2019, 59-68.

[20] Usp. Jon M. KLEINBERG, Sendhil MULLAINATHAN, Manish RAGHAVAN, Human Decisions and Machine Predictions, *The quarterly journal of economics*, 133 (2018) 1, 237-293.

[21] Usp. Solon BAROCAS, Andrew SELBST, Big Data's Disparate Impact, *California Law Review*, 104 (2016) 671-732, https://ssrn.com/abstract=2477899.

[22] Usp. Alexandra CHOULDECHOVA, Fair Prediction with Disparate Impact. A Study of Bias in Recidivism Prediction Instruments, *Big data*, 5 (2017) 2, 153-163, doi:10.1089/big.2016.0047.

[23] Usp. Sam CORBETT-DAVIS i dr., The Measure and Mismeasure of Fairness. The Journal of Machine Learning Research, 24 (2023) 1, 14730-14846.

[24] Usp. Simon CATON, Christian HAAS, Fairness in Machine Learning. A Survey, *ACM Computing Surveys*, 56 (2020) 7, 1-38.

izračunavanju i izražavanju pravednosti s obzirom na to da uzroci nepravednosti nadilaze baze podataka i matematičke izračune.

No, iako postoje i razne definicije pravednosti, većina znanstvenika smatra da pravednost zahtijeva nepristranu raspodjelu vrijednosti bez osvrtanja na pojedine karakteristike pojedinca ili njihovu situaciju koja se smatra nevažnom.[25] Drugim riječima, predviđanja koja donosi sustav UI ne bi smjela proizvesti nepravedne, diskriminirajuće ili disparatne posljedice.[26] Zbog toga postizanje pravednosti zahtijeva sveobuhvatno razumijevanje potencijalnih uzroka pristranosti u umjetnoj inteligenciji i razvoj strategija za ublažavanje pristranosti.[27]

Zbog sve veće primjene UI u brojnim područjima ljudske djelatnosti, trenutno se razvija više etičkih i pravnih smjernica za brojna područja, od kojih je zdravstvo – s obzirom na osjetljivost, složenost definicija, obveza i zahtjeva –posebno važno područje.[28] Akt o UI, jedan od vodećih pravnih dokumenata u području regulacije UI, izdvaja sustave UI koji se primjenjuju u zdravstvu kao visokorizične koji podliježu strogim regulacijama koje je potrebno zadovoljiti prije negoli se takav sustav UI smije upotrebljavati u zdravstvenom kontekstu. S obzirom na to da sustavi UI mogu brzo analizirati veliku količinu podataka te identificirati obrasce i trendove, to omogućuje razvoj etičkih i pravnih smjernica utemeljenih na dokazima u stvarnom vremenu, što značajno olakšava razmjenu informacija s vodećim stručnjacima radi razvoja etičkih implikacija.[29]

Brojne međunarodne organizacije razvijaju okvire za regulaciju UI u specifičnim zdravstvenim situacijama, poput Američke agencije za hranu i lijekove koja izrađuje smjernice za kritičku procjenu stvarnih primjena UI u medicini, a već je objavila okvir za ulogu UI i strojnog učenja u softveru kao medicinskom uređaju.[30] Također, Europska komisija je poduzela multidisciplinarne napore za jačanje povjerenja u primjenu UI,[31] dok Europska agencija za lijekove regula-

[25] Usp. Paula BODDINGTON, *AI Ethics. A Textbook*, Singapore, Springer, 2023, 714.

[26] Usp. Don Donghee SHIN, Yong Jin PARK, Role of fairness, accountability, and transparency in algorithmic affordance, *Computers in Human Behavior*, 98 (2019) 277-284.

[27] Usp. María Agustina RICCI LARA, Rodrigo ECHEVESTE, Enzo FERRANTE, Addressing fairness in artificial intelligence for medical imaging, *Nature communications*, 13 (2022) 1, 1-6, doi:10.1038/s41467-022-32186-3.

[28] Usp. Felix BUSCH i dr., Navigating the European Union Artificial Intelligence Act for Healthcare, *NPJ digital medicine*, 7 (2024) 1, 1-6, doi:10.1038/s41746-024-01213-6.

[29] Usp. Shuroug ALOWAIS i dr., Revolutionizing healthcare. The role of artificial intelligence in clinical practice, *BMC medical education*, 23 (2023) 1, 1-15, doi:10.1186/s12909-023-04698-z.

[30] Usp. AMERIČKA AGENCIJA ZA HRANU I LIJEKOVE, *Umjetna inteligencija i strojno učenje u softveru kao medicinskom uređaju*, https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (14.01.2025).

[31] Usp. EUROPSKA KOMISIJA, *Bijela knjiga o umjetnoj inteligenciji. Europski pristup izvrsnosti i povjerenju*, https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (14.01.2025).

ciju UI smatra strateškim prioritetom.[32] Uloga UI u uspostavljanju smjernica je pružanje uvida i preporuka temeljenih na podatcima učinkovitijeg donošenja odluka, boljih rezultata liječenja i smanjenja troškova.

Budućnost zdravstvenih sustava visokih performansi vjerojatno će se oslanjati na sinergiju temeljenu na interakciji čovjeka i umjetne inteligencije.[33] UI ima velik potencijal za razvoj i korištenje u zemljama s visokim nacionalnim bruto dohotkom u mnogim različitim područjima liječenja, uključujući radiologiju, patologiju, dermatologiju, intenzivnu njegu, oftalmologiju, kardiologiju.[34] Glavnih razlog za sve veću primjenu sustava UI jest što oni omogućavaju značajnu moć predviđanja, dijagnostičku preciznost i više prilagođenih recepata za lijekove. Iako je potrebno provesti proces detaljnog potvrđivanja kliničke primjene i učinkovitosti sustava UI, poznato je da neki algoritmi danas nadmašuju ljudske stručnjake (liječnike) u otkrivanju bolesti, brzini tumačenja, procjeni rizika za ponovnu pojavnost bolesti i mortalitet i klasifikaciju rendgenskih nalaza.[35] Budući da se digitalne tehnologije sve više koriste u kontekstu zdravstvene skrbi, povećava se i rizik od nanošenja štete pojedincima i društvenim skupinama. Stoga pravednost postaje jedan od ključnih principa za zdravstvenu skrb koja – u skladu s algoritamskom pravednošću – pokušava maksimalno nepristrano i objektivno pristupiti svim pacijentima.

Razni izazovi, poput terminoloških nejasnoća, pristranosti, nepravednosti, nedostatka objašnjivosti i povjerenja, čine pukotine prilikom primjene UI u zdravstvenim sustavima. Nepristran i objektivan pristup svim pacijentima trebao bi biti cilj kojeg želi doseći zdravstveno osoblje prilikom razmatranja odluka poduprtih algoritamskim odlučivanjem koje uključuje neki od sustava pravednosti.

2. Terminološka razdioba pravednosti u zdravstvenoj skrbi

S obzirom na izazov terminoloških nejasnoća, pravednost može biti višeznačan pojam koji ovisi o različitim kontekstima unutar kojih pokušavamo definirati pravednost, kao što su: pravo, društvene znanosti, STEM područja ili filozofija. Tako bi, na primjer, definicija pravednosti u kontekstu društvenih znanosti bila promatrana u svjetlu društvenih odnosa, dinamike moći, insti-

[32] Usp. EUROPSKA AGENCIJA ZA LIJEKOVE, *Regulatorna znanstvena strategija do 2025.*, https://www.ema.europa.eu/en/about-us/how-we-work/regulatory-science-strategy#regulatory-science-strategy-to-2025-section (14.01.2025).

[33] Usp. Eric J. TOPOL, High-performance medicine: the convergence of human and artificial intelligence, *Nature medicine*, 25 (2019) 1, 44-56, doi:10.1038/s41591-018-0300-7.

[34] Usp. THE LANCET PUBLIC HEALTH, Next generation public health: towards precision and fairness, *The Lancet. Public health*, 4 (2019) 5, 1, doi:10.1016/S2468-2667(19)30064-7.

[35] Usp. Jon RUEDA i dr., »Just« accuracy? Procedural fairness demands explainability in AI-based medical resource allocations, *AI & society*, 39 (2022) 1-12, doi:10.1007/s00146-022-01614-9.

tucija i tržišta, dok bi u kontekstu filozofije definicija pravednosti počivala na ideji da »ono što je pravedno je ujedno i moralno«.[36] Jedna od čestih definicija pravednosti jest kvaliteta postupanja prema ljudima na jednak način ili na način koji je ispravan ili razuman,[37] a za sustav UI se kaže da je pošten ako ne diskriminira pojedinca ili grupu.[38] No, usprkos tim definicijama, postoje brojni izazovi u pristupu i stavovima kada pokušavamo konceptualizirati pravednost u kontekstu zdravstvene skrbi.

Specifični izazovi terminološke razdiobe tiču se matematičkih pokušaja formulacije pravednosti i eventualnih preferencija koje sustav UI može imati za pojedini model. Matematička kvantifikacija pravednosti ima smisla budući da pravednost ima tendenciju odgovarati nekoj vrsti kriterija, kao što su jednaka ili pravedna raspodjela, ili zastupljenost pogreške za pojedini problem i/ili zadatak. Do sada se pravednost pokušala definirati brojnim matematičkim definicijama koje uključuju demografski paritet,[39] disparatni učinak,[40] jednakost prilika[41] i kalibraciju.[42] Svaka navedena matematička formulacija privlačna je na svoj način, no pokazalo se da su navedene definicije međusobno nekompatibilne te se ne mogu istovremeno zadovoljiti.[43] Neki znanstvenici su bezuspješno pokušali ronaći kompromise između različitih protuslovnih formulacija pravednosti.[44] No drugi autori potvrđuju da je pomirenje tih protuslovlja neizvedivo,[45] zbog čega se predlaže drukčiji pristup i definicija pravednosti, ona ovisnu o

[36] Deirdre MULLIGAN i dr., This thing called fairness. Disciplinary confusion realizing a value in technology, *ACM Human-Computer Interaction*, 3 (2019) 1-36, https://doi.org/10.1145/3359221.

[37] Usp. CAMBRIDGE DICTIONARY, https://dictionary.cambridge.org/dictionary/english/fairness (14.01.2025).

[38] Usp. Solon BAROCAS, Moritz HARDT, Arvind NARAYANAN, Fairness and machine learning: limitations and opportunities, Cambridge, *The MIT Press*, 2023, 1-340.

[39] Usp. Cynthia DWORK i dr., Fairness through awareness, u: Shafi Goldwasser (ur.), *Proceedings of the Innovations in Theoretical Computer Science Conference*, New York, Association for Computing Machinery, 2012, 214-226.

[40] Usp. Muhammad Bilal ZAFAR i dr., Fairness Beyond Disparate Treatment & Disparate Impact. Learning Classification without Disparate Mistreatment, u: Rick Barrett i dr. (ur.), *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017, 1171-1180.

[41] Usp. Moritz HARDT, Eric PRICE, Nati SREBRO, Equality of Opportunity in Supervised Learning, *Neural Information Processing Systems*, (2016) 1-9.

[42] Usp. Kleinberg i dr., *Inherent Trade-Offs…*, 2.

[43] Usp. Alexandra CHOULDECHOVA, Fair prediction with disparate impact. A study of bias in recidivism prediction instruments, *Big data*, 5 (2016) 2, 153-163.

[44] Usp. Sam CORBETT-DAVIES i dr., Algorithmic Decision Making and the Cost of Fairness, u: Stan Matwin i dr. (ur.), *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, Association for Computing Machinery, 2017, 797-806.

[45] Usp. Megha SRIVASTAVA, Hoda HEIDARI, Andreas KRAUSE, »Mathematical Notions vs. Human Perception of Fairness. A Descriptive Approach to Fairness for Machine Learning«, u: Ankur Teredesai i dr. (ur.), *Proceedings of the 25th ACM SIGKDD International Conference on*

specifičnom kontekstu zdravstvene skrbi. Dakle, umjesto da se zahtijeva da svi pojmovi pravednosti – makar djelomično – vrijede u isto vrijeme, smatra se da je bolji pristup određivanje najprikladnijeg pojma pravednosti za specifično zdravstvenu domenu o kojoj se nešto više govori u nastavku.

Druga vrsta terminoloških izazova tiče se specifično tehničke perspektive unutar koje se razvijaju različiti modeli pravednosti. Rijetko kada su se različite definicije i pojmovi relevantni za pravednost uzimali u obzir prije razvoja sustava UI. Čak i u situacijama kada se razmotre različite definicije ili pristupi pravednosti, čak ni tada ne postoji »pravedan« odgovor za pojedinačan sustav UI. Osim toga, u proces strojnog učenja (od razvoja skupa podataka preko razvoja algoritma i upotrebe sustava UI) uključeni su razni akteri koji mogu imati različita tumačenja i razumijevanja pravednosti, što može dodatno doprinijeti razvoju terminoloških nejasnoća.[46] Također, specifičan tehnički izazov tiče se postavljanja granica, a postavi li se previše ograničenja za detekciju pravednosti sustavu UI predviđanja će imati manju preciznost.[47] Uz to, posebno zanimljiv tehnički izazov formalizacije pravednosti odnosi se na neprozirnosti samog sustava UI. Novi uzrok nepravednosti može proizlaziti iz toga što korisnik sustava UI ne može razumjeti predviđanja koja su došla iz sustava UI te je zbog toga važno raditi na transparentnosti i objašnjivosti da bi se omogućilo okolnosti u kojima korisnik razumije predviđanja i u kojima je razvoj pravednosti moguć.[48]

Treća vrsta terminoloških prepreka tiče se pitanja zdravstvenih nejednakosti (engl. *disparities*) koje nastaju kao rezultat povijesnih i trenutnih socioekonomskih nejednakosti te utječu na kvalitetu zdravstvene skrbi među pojedinim skupinama pacijenata. Zdravstvene nejednakosti se pripisuju širokom rasponu uzroka rizika, uključujući indeks tjelesne mase, obrazovanje, vrstu osiguranja, genetiku. Većina ovih uzroka grupirana je u pet sljedećih domena društvenih odrednica zdravlja: ekonomska stabilnost, pristup obrazovanju i njegova kvaliteta, pristup zdravstvenoj skrbi i njezina kvaliteta, susjedstvo i izgrađeno okruženje, te kontekst društva i lokalne zajednice.[49] Važno je istaknuti da se ti čimbenici pripisuju različitim zdravstvenim ishodima i nepovjerenju u zdravstveni sustav.[50] Razumijevanje ovih nejednakosti može pomoći pri usmjeravanju razvoja zdravstvene skrbi i sprječavanja daljnjeg povećanja nejednakosti ili smanjenja kvalitete zdravstvene skrbi, a zbog pojave nejednakosti u zdravstvu

Knowledge Discovery & Data Mining, New York, Association for Computing Machinery, 2019, 2459-2468.

[46] Usp. Genevieve SMITH, *What does »fairness« mean for machine learning systems?*, https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf, (14.01.2025).

[47] Usp. Mulligan i dr., *This thing called fairness…*, 15.

[48] Usp. *isto*, 20.

[49] Usp. Richard J. CHEN i dr., Algorithmic fairness in artificial intelligence for medicine and healthcare, *Nature biomedical engineering*, 7 (2023) 6, 719-742, doi:10.1038/s41551-023-01056-8.

[50] Usp. Ana I. BALSA i dr., Clinical Uncertainty and Healthcare Disparities, *American Journal of Law & Medicine*, 29 (2003) 2-3, 203-219.

uzrokovane sustavima UI procjena i ublažavanje štete postali su središnja motivacija za proučavanje pravednosti u strojnom učenju.[51]

Nejednakosti mogu nastati i u ishodima i u praksi zdravstvene skrbi zbog varijacija u prevalenciji bolesti, kvaliteti usluga i različitih uređaja koji se koriste za isti medicinski postupak u različitim regijama i bolnicama (kao što su to, na primjer, različiti uređaji za analizu krvi).[52] Sustavi UI iz jedne bolnice se možda ne mogu prilagoditi sustavu UI drugih bolnica zbog posebnih metoda korištenih za prikupljanje, obradu i organiziranje podataka koji mogu imati nenamjerno kodirane pristranosti specifične za bolnicu u kojoj se podaci obrađuju.[53] Osim toga, nemaju sve skupine ljudskih populacija dugu povijest prikupljanja podataka u postojećim skupovima biomedicinskih podataka.[54]

Radi uvećanja količine dostupnih podataka objedinjuju se skupovi podataka iz više bolnica. No, budući da različiti bolnički centri mogu imati različite količine dostupnih podataka u obuci, sustavi UI mogu akumulirati pristranosti specifične za tu bolnicu tijekom obuke. Kada se te pristranosti odraze u predviđanjima sustava UI, tada neke bolnice mogu biti izolirane zbog lošijih ishoda,[55] što može povećati nejednakost među bolnicama i obeshrabriti te bolnice za implementaciju tehnologija temeljenih na sustavima UI.[56]

Jedno od potencijalnih rješenja za sve navedene terminološke izazove definiranja pravednosti u zdravstvenom kontekstu neki autori su pokušali razriješiti primjenom definicije distributivne pravednosti s ova tri elementa: jednak ishod, jednak učinak i jednaka raspodjela. Ova tri elementa osiguravaju jednake rezultate za sve društvene skupine, s naglaskom na jednaku preciznost, osjetljivost, specifičnost i pozitivnu prediktivnu vrijednost.[57] Međutim, ovakav pristup je kritiziran jer se previše oslanja na demografski paritet i matematičko shvaćanje pravednosti, što je već prije spomenuto.

Ipak, s obzirom na to da pacijenti nisu puki biološki organizmi, već ljudska bića s općim i individualiziranim potrebama, željama, ranjivostima i vrijed-

[51] Usp. Chen, *Algorithmic fairness in artificial intelligence*…, 721.

[52] Usp. Manhal ALI, Reza SALEHNEJAD, Mohaimen MANSUR, Hospital heterogeneity: what drives the quality of health care, *The European journal of health economics. HEPAC: health economics in prevention and care*, 19 (2018) 3, 385-408, doi:10.1007/s10198-017-0891-9.

[53] Usp. Jenny YANG, Andrew A. S. SOLTAN, David A. CLIFTON, Machine learning generalizability across healthcare settings. Insights from multi-site COVID-19 screening, *NPJ digital medicine*, 5 (2022) 1, 1-8, doi:10.1038/s41746-022-00614-9.

[54] Usp. Natalia NORORI i dr., Addressing bias in big data and AI for health care. A call for open science, *Patterns*, 2 (2021) 10, 1-9, doi:10.1016/j.patter.2021.100347.

[55] Usp. Jenny YANG i dr., Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning, *Nature machine intelligence*, 5 (2023) 8, 884-894, doi:10.1038/s42256-023-00697-3.

[56] Usp. Yang i dr., *Algorithmic fairness and bias mitigation*…, 2.

[57] Usp. Benedetta GIOVANOLA and Simona TIRIBELLI, Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms, *AI & society*, 38 (2023) 2, 549-563, doi:10.1007/s00146-022-01455-6.

nostima[58] važno je u kontekst pravednosti uzeti sve navedene kategorije kao i uključiti jedinstven odnos između zdravstvenog osoblja i pacijenta prožet posebnim vrijednostima i dužnostima.[59] Općenito se smatra da ovaj odnos zahtijeva pristup usmjeren na pacijenta koji poštuje autonomiju pacijenta i promiče informirani pristanak. Stoga se sljedeća definicija pravednosti može prihvatiti kao relevantna za kontekst zdravstva: pravednost u zdravstvenoj skrbi je višedimenzionalan koncept koji uključuje pravednu raspodjelu resursa, mogućnosti i ishode među različitim populacijama pacijenata.[60] U definiciju pravednosti trebali bi biti integrirani postojeći zakonski propisi svake zemlje i koncept društvene pravednosti da bi se izbjegla uska tehnička rješenja.[61]

3. (Ne)pristrana umjetna inteligencija u zdravstvenoj skrbi

U javnom prostoru se može čuti da se sustavi UI u procesu odlučivanja mogu koristiti kao alati za postizanje pravednosti i jednakosti zahvaljujući algoritamskoj sposobnosti prevladavanja kognitivnih ograničenja i društvenih predrasuda proizašlih iz ljudske prirode, što omogućuje objektivnije i pravednije odluke.[62] No, brojni primjeri pokazuju da to zapravo nije uvijek slučaj. Slike, tekst i numeričke vrijednosti kao ulazni podaci na temelju kojih se sustavi UI razvijaju bogati su pristranostima. Do sada je dokazano kako u situacijama kada u donošenju odluka sudjeluje UI, takve odluke mogu biti pristrane i diskriminirajuće[63] što u zdravstvu može biti posebno izazovno s obzirom na to da je riječ o situacijama visokog rizika i o osjetljivim podatcima.[64] Neki autori tvrde da je nemoguće kreirati sustav UI bez pristranosti te je potrebno voditi računa o pristranostima uključenima u sustave UI.[65] Neki autori tvrde da je problem pristranosti kompleksan društveni problem kojeg nije lako tehnički

[58] Usp. Paul RAMSEY, *The Patient as Person: Explorations in Medical Ethics*, London, Yale University Press, 2002, 268.

[59] Usp. Thomas P. QUINN i dr., Trust and medical AI. The challenges we face and the expertise needed to overcome them, *Journal of the American Medical Informatics Association*, 28 (2021) 4, 890-894, doi:10.1093/jamia/ocaa268.

[60] Usp. Michael MARMOT, Ruth BELL, Fair Society, Healthy Lives, *Public health*, 126 (2012) 4-10, https://doi.org/10.1016/j.puhe.2012.05.014.

[61] Usp. Xiaomeng WANG, Yishi ZHANG, Ruilin ZHU, A brief review on algorithmic fairness, *MSE*, 1 (2022) 7, 1-13, https://doi.org/10.1007/s44176-022-00006-z.

[62] Usp. Ben GREEN, Escaping the Impossibility of Fairness. From Formal to Substantive Algorithmic Fairness, *Philosophy & Technology*, 35 (2022) 1-32, https://doi.org/10.1007/s13347-022-00584-6.

[63] Usp. Catherine STINSON, Algorithms are not neutral. Bias in collaborative filtering, *AI and ethics*, 2 (2022) 4, 763-770, doi:10.1007/s43681-022-00136-w.

[64] Usp. Marvin van BEKKUM, Frederik J. Zuiderveen BORGESIUS, Using sensitive data to prevent discrimination by artificial intelligence. Does the GDPR need a new exception?, *Computer Law and Security Review*, 48 (2023) 1-12.

[65] Usp. Paulo CARVÃO, *Can AI Be Fair and Unbiased?* https://www.sir.advancedleadership.harvard.edu/articles/can-ai-be-fair-and-unbiased, (14.01.2025).

riješiti, niti imamo dovoljno jasno razumijevanje o radu specifičnih sustava UI kao što su veliki jezični modeli.[66]

Sustavi UI mogu razviti pristranost, na primjer, pri oslanjanju na povijesne podatke o uhićenjima,[67] kao što je to bilo na primjeru prediktivnog sustava UI za ponavljanje kaznenih djela pod nazivom COMPAS. Iako se očekivalo da će predviđanja biti pravedna i nepristrana, to se nije dogodilo. COMPAS je bio pristran i davao lažno pozitivna predviđanja za optuženike crne boje kože. Lažno ih je označavao kao buduće počinitelje kaznenih djela gotovo dvostruko češće nego bijele optuženike. Slučaj je postao poznat nakon članka ProPublice.[68]

Slični primjeri nepravednih i pristranih predviđanja događali su se u kliničkim okruženjima gdje su istraživači ustanovili slično – da sustavi UI stvaraju pristrana i diskriminirajuća predviđanja na temelju rasne, etničke ili nacionalne pripadnosti ili drugih nejednakosti. Posljedice pristranosti i diskriminacije prilikom predviđanja od strane sustava UI su se do sada proučavale u zdravstvenoj skrbi u različitim domenama poput dijagnosticiranja, izbora liječenja i učinkovitosti zdravstvenog sustava[69] ili obrade prirodnog jezika i dubokog učenja.[70] Također, medicinska oprema može stvarati netočne podatke kao što se to dogodilo u primjerima pogrešnog očitanja krvnog tlaka[71] ili pogrešnom očitanju zasićenosti kisika pulsnog oksimetra kod pacijenta s crnom bojom kože.[72] Razina zasićenosti kisikom u krvi mjeri se pulsnim oksimetrom slanjem infracrvenog svjetla kroz kožu. Boja pacijentove kože utječe na mjerenja pulsnog oksimetra jer uređaj krivo procjenjuje razine zasićenja kisikom kod pacijenata koji nisu bijele rase. Na temelju toga mjerenja zaključeno je da pacijenti crne rase imaju tri puta veću vjerojatnost da obole od akutne hipoksemije, a to pulsni oksimetar kod pacijenata crne boje kože ne otkriva na isti način kao kod pacijenata bijele boje kože. Pristranost u zdravstvenoj skrbi mogu započeti kliničkim mjerenjima koji uzrokuju pogrešne medicinske odluke za skupine

[66] Usp. Melissa HEIKKILÄ, *Why it's impossible to build an unbiased AI language model*, https://www.technologyreview.com/2023/08/08/1077403/why-its-impossible-to-build-an-unbiased-ai-language-model/, (14.01.2025).

[67] Usp. Kleinberg i dr., *Inherent Trade-Offs…*, 1.

[68] Usp. Anne WASHINGTON, How to Argue with an Algorithm. Lessons from the COMPAS ProPublica Debate, *The Colorado Technology Law Journal*, 17 (2019) 1, 1-37.

[69] Usp. Rajkomar i dr., *Ensuring Fairness in Machine Learning…*, 1.

[70] Usp. Ninareh MEHRABI i dr., A Survey on Bias and Fairness in Machine Learning, u: Albert Zomaya (ur.), *ACM Computing Surveys*, New York, Association for Computing Machinery, 2019, 1-35.

[71] Usp. Junichi ISHIGAMI i dr., Effects of Cuff Size on the Accuracy of Blood Pressure Readings. The Cuff (SZ) Randomized Crossover Trial, *JAMA internal medicine*, 183 (2023) 10, 1061-1068, doi:10.1001/jamainternmed.2023.3264.

[72] Usp. Ana M. CABANAS, Pilar MARTÍN-ESCUDERO, Kirk H. SHELLEY, Improving pulse oximetry accuracy in dark-skinned patients: technical aspects and current regulations, *British journal of anaesthesia*, 131 (2023) 4, 640-644, doi:10.1016/j.bja.2023.07.005.

pacijenata i mogu se dalje krivo usmjeriti primjenom predviđanja sustava UI u zdravstvu.[73]

Uz to, sustavi UI koriste se i u jedinicama intenzivnog liječenja te prilikom praćenja pacijenata preporučuje brzu medicinsku intervenciju za visokorizične pacijente. Međutim, ako je baza podataka testirana samo na jednoj skupini ljudi, predviđanja sustava UI će biti neprecizna i netočna za skupine ljudi na kojima nije testirana. To dovodi do lažno pozitivnih predviđanja, uzrokujući da liječnici ignoriraju znakove upozorenja i smanje povjerenje u predviđanja sustava UI što u sklopu jedinice intenzivnog liječenja može uzrokovati potencijalno i smrtonosne zdravstvene ishode.[74]

U području kirurgije autori Kiyaseeh, Laca i ostali su razvili sustav UI pod nazivom SAIS, za procjenu vještina kirurga, koji je pokazao pristranost prilikom procjene vještine kirurga u različitim kirurškim zahvatima.[75] Iako je SAIS mogao i pouzdano procijeniti kirurški učinak, događale su se i situacije u kojima je pristranošću došlo do značajnih pogrešaka u predviđanju. SAIS sustav je pogrešno smanjio stupanj vještine kirurške izvedbe predviđajući da će vještina kirurga biti niže kvalitete nego što je zapravo bila. S druge strane, SAIS je i pogrešno unaprijedio kvalitetu kirurške izvedbe predviđajući da će neka kirurška vještina biti kvalitetnija nego što je bila.[76]

Poznato je i da su u razvoj i primjenu sustava UI uključene različite pristranosti koje mogu negativno utjecati na ispravnost predviđanja pogrešnim iskrivljavanjem rezultata sustava UI.[77] Važno je imati na umu da iskrivljavanje rezultata može nastati u bilo kojem trenutku obrade ili čuvanja podataka te se zbog toga nije moguće slažiti s Groteom i Keelingom da je jedino važno voditi računa o pravednosti konačnih odluka,[78] nego je važno imati na umu sve ishode predviđanja.

Istraživanja iz područja algoritamske pravednosti naglašavaju da nesavršeni podaci mogu odvesti na krivi put, ali i da sami sustavi UI nisu vrijednosno neutralni. Razlozi za nepravednost mogu biti mnogi i najčešće mogu nastati iz dva glavna izvora podataka koji se koriste za obuku algoritma (pristranost podata-

[73] Usp. Sylvia E. K. SUDAT i dr., Racial Disparities in Pulse Oximeter Device Inaccuracy and Estimated Clinical Impact on COVID-19 Treatment Course, *American journal of epidemiology*, 192 (2023) 5, 703-713, doi:10.1093/aje/kwac164.

[74] Usp. Angier ALLEN i dr., A Racially Unbiased, Machine Learning Approach to Prediction of Mortality. Algorithm Development Study, *JMIR public health and surveillance*, 6 (2020) 4, 1-9, doi:10.2196/22400.

[75] Usp. Dani KIYASSEH i dr., Human visual explanations mitigate bias in AI-based assessment of surgeon skills, *NPJ digital medicine*, 6 (2023) 1, 1-12, doi:10.1038/s41746-023-00766-2.

[76] Usp. Mirja MITTERMAIER, Marium M. RAZA, Joseph C. KVEDAR, Bias in AI-based models for medical applications. Challenges and mitigation strategies, *NPJ digital medicine*, 6 (2023) 1, 1-3, doi:10.1038/s41746-023-00858-z.

[77] Usp. Bærøe i dr., *Can medical algorithms be fair?...*, 1.

[78] Usp. Thomas GROTE, Geoff KEELING, Enabling Fairness in Healthcare Through Machine Learning, *Ethics and information technology*, 24 (2022) 3, 1-13, doi:10.1007/s10676-022-09658-7.

ka) i inherentnog dizajna ili mehanizama učenja samog algoritma (algoritamska pristranost).[79] No česti su i ostali izvori nepravednosti poput pristranosti koja proizlazi iz povijesnih činjenica, pristranost koja proizlazi iz algoritamskih ciljeva, neuravnotežen skup podataka, nepotpuni podaci ili neprozirnost algoritma.[80] U kontekstu zdravstvene skrbi mogu se pojaviti dodatne pristranosti zbog složene prirode ljudskih interakcija i procesa donošenja odluka. Neki autori navode da se u specifičnoj domeni zdravstvene skrbi pristranosti mogu svrstati u sljedeće četiri kategorije:

1) pristranosti u sustavima UI ovisne o dizajnu modela (pristranosti oznake i pristranosti skupine),
2) pristranosti u sustavima UI ovisne o podacima o obuci (pristranost manjine, pristranost podataka koji nedostaju, pristranost informativnosti),
3) pristranosti u sustavima UI uzrokovane interakcijama sustava UI s liječnicima (pristranost automatizacije, petlje povratnih informacija),
4) pristranosti u sustavima UI uzrokovane interakcijama sustava UI s pacijentima (pristranost prema privilegijama, informirano nepovjerenje).[81]

Navedeni slučajevi pristranosti i nepravednosti postavljaju središnje pitanje o upotrebi sustava UI u zdravstvu: Kako liječnici mogu biti sigurni da donose pravednu i nepristranu odluku uz pomoć sustava UI? I na što bismo točno trebali usmjeriti svoju pažnju?

Da bi se smanjio utjecaj pristranosti, moguće je koristiti različite oblike strategija za otkrivanje i ublažavanje nepravednosti i pristranosti koje su ključne za poboljšanje ishoda zdravstvene skrbi. Jedan od dostupnih alata koji može pomoći u smanjenju utjecaja pristranosti je i OUI koji je već prije spomenut.

4. Prema ostvarenju pravednosti u zdravstvenoj skrbi

Trenutno postoje razni alati koji znanstvenicima i istraživačima pokušavaju pomoći u identifikaciji podataka i ublažavanju nepravednosti i pristranosti u sustavima UI kao što su »AI Fairness 360 Toolkit« u vlasništvu tvrtke IBM, »What-If Tool« u vlasništvu tvrtke Google, »fairlean.py« u vlasništvu tvrtke Microsoft i »Fairness Flow« u vlasništvu tvrtke Facebook.[82] Ovi alati pružaju tehnička rješenja, metriku i algoritme za poboljšanje preciznosti i pouzdanosti sustava UI.

[79] Usp. Ueda i dr., *Fairness of artificial intelligence in healthcare…*, 4.
[80] Usp. Ravi B. PARIKH, Stephanie TEEPLE, Amol S. NAVATHE, Addressing Bias in Artificial Intelligence in Health Care, *JAMA*, 322 (2019) 24, 2377-2378, doi:10.1001/jama.2019.18058.
[81] Usp. Rajkomar i dr., *Ensuring Fairness in Machine Learning…*, 8-9.
[82] Usp. Smith, *What does »fairness« mean...*, 4.

No sva napredna tehnička rješenja koja pokušavaju bez mogućnosti objašnjenja riješiti kompleksne etičko-epistemološke izazove ne nude odgovore na pitanja o profesionalnoj odgovornosti. Profesionalna odgovornost podrazumijeva da je opravdano tražiti od zdravstvenog djelatnika da objasni svoje postupke te jasno artikulira i obrazloži odluke koje je donio. Pružanje takvog objašnjenja stvara povjerenje u proces koji je doveo do odluke i povjerenje da je zdravstveni radnik zadužen za proces postupio pravedno i razumno.[83] Istina je, kao što neki autori ističu, da nedostatak objašnjivosti u zdravstvu nije neuobičajen, a ponekad je gotovo nemoguće precizno rekonstruirati objašnjenje na kojem se temelji klinička prosudba zbog nedostatka znanja o uzročnim mehanizmima kroz koje intervencije djeluju.[84] Drugi autori tvrde da pacijenti rutinski oslanjaju na farmakološke tretmane za koje ne znaju kako djeluju, ali znaju da djeluju. U tim slučajevima objašnjivost može biti korisna, ali znanstvenici se ne slažu u tomu da bi načelo objašnjivosti trebalo biti uvijek potpuno ispunjeno.[85]

Međutim, objašnjivost je još uvijek posebno važna u situacijama koje zahtijevaju informirani pristanak, a nedavna pravna regulacija *Akta o umjetnoj inteligenciji* dodatno je stavila naglasak na važnost objašnjivosti, transparentnosti i ljudskog nadzora.[86] Štoviše, neki autori smatraju da sustavi UI bez objašnjivog sustava potpore kliničkom odlučivanju ugrožavaju temeljne etičke vrijednosti u medicini (uključujući tradicionalna bioetička načela autonomije, dobročinstva, neškodljivosti i pravde) i mogu čak uzrokovati štetne posljedice za pojedince, kao i za javno zdravlje.[87]

Brojni autori navode da sustavima UI često nedostaje transparentnost, što otežava objašnjenje njihovih predviđanja i narušava povjerenje.[88] Taj nedostatak transparentnosti može uzrokovati lošije ishode za skupine koje nisu dovoljno zastupljene, kao što je prije navedeno.[89] Uključivanjem transparentnosti u više točaka razvoja, testiranja i faza implementacije primjene sustava

[83] Usp. Bærøe i dr., *Can medical algorithms be fair?...*, 5.

[84] Usp. Alex John LONDON, Artificial Intelligence and Black-Box Medical Decisions. Accuracy versus Explainability, *The Hastings Center report*, 49 (2019) 1, 15-21, doi:10.1002/hast.973.

[85] Usp. Scott ROBBINS, A misdirected principle with a catch. Explicability for AI, *Minds Mach*, 29 (2019) 4, 495-514, https://doi.org/10.1007/ s11023-019-09509-3.

[86] Usp. Cecilia PANIGUTTI i dr., The role of explainable AI in the context of the AI Act, *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, New York, Association for Computing Machinery, 2023, 1139-1150, https://doi.org/10.1145/3593013.3594069.

[87] Usp. Julia AMANN i dr., Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, *BMC medical informatics and decision making*, 20 (2020) 1, 1-9, doi:10.1186/s12911-020-01332-6.

[88] Usp. Seyedeh Neelufar PAYROVNAZIRI i dr., Explainable artificial intelligence models using real-world electronic health record data. A systematic scoping review, *Journal of the American Medical Informatics Association*, 27 (2020) 7, 1173-1185, doi:10.1093/jamia/ocaa053.

[89] Usp. Thomas GROTE, Philipp BERENS, On the ethics of algorithmic decision-making in healthcare, *Journal of medical ethics*, 46 (2020) 3, 205-211, doi:10.1136/medethics-2019-105586.

UI u zdravstvu[90] može se napraviti važan korak za više pravednosti i smanjenje utjecaja pristranosti. Ako se proces primjene UI učini transparentnijim, bit će lakše prepoznati i ukloniti pristranost. Koncepti pravednosti i transparentnosti uvelike ovise o tome kako se obrađuju i analiziraju podatci, kako se koristi UI i što se uključuje u dizajn samog procesa. Transparentnost je jedno od načela o kojem se najviše govori u raspravi o etici UI i postaje jedna od značajki koja se može definirati kao »potraga za izravnim razumijevanjem mehanizma po kojem model funkcionira«.[91] Transparentnost i objašnjivost sustava UI ključni su za razvoj koncepta povjerenja. Od iznimne je važnosti raditi na pravednoj primjeni UI-e u zdravstvu da bi liječnici i pacijenti mogli vratiti povjerenje u primjenu UI u zdravstvu.

Objašnjivost u zdravstvenoj skrbi je istraživana u brojnim radovima, a detaljno su obrađene i neke tehničke ograničenosti OUI[92] kao i u zdravstvenoj skrbi.[93] Iako se o objašnjivosti može široko raspravljati, važno je napomenuti da je nedostatak objašnjivosti izrazito važan za pravednost s obzirom na to da je bez objašnjivosti teško utvrditi je li sustav UI uzeo u obzir etički relevantne varijable. Izostanak objašnjivosti ugrožava odgovornost u slučaju etičkih pristranosti ili negativnih posljedica za primatelje, smanjuje povjerenje u sustav UI i potencijalno ugrožava javnu podršku njezinoj primjeni u zdravstvu.[94] Mnogi autori naglašavaju da je za razvoj povjerenja u sustave UI potrebna objašnjivost, primjerice: »Da bi ljudi vjerovali metodama crne kutije, potrebna nam je objašnjivost…«;[95] »Potreba za objašnjivom umjetnom inteligencijom motivirana je uglavnom (...) potrebom za povjerenjem…«;[96] »postoji potreba za objašnjenjem (...) da bi korisnici i donositelji odluka mogli razviti odgovarajuće povjerenje«;[97]

[90] Usp. Melissa D. McCRADDEN i dr., Ethical limitations of algorithmic fairness solutions in health care machine learning, *The Lancet. Digital health*, 2 (2020) 5, 221-223, doi:10.1016/S2589-7500(20)30065-0.

[91] Tugba AKINCI D'ANTONOLI, Ethical considerations for artificial intelligence. An overview of the current radiology landscape, *Diagnostic and interventional radiology*, 26 (2020) 5, 504-511, doi:10.5152/dir.2020.19279.

[92] Usp. Sebastian BORDT i dr., Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts, *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, New York, Association for Computing Machinery, 2022, 891-905, https://doi.org/10.1145/3531146.3533153.

[93] Usp. Tim HULSEN, Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare, *AI*, *4*, (3), (2023), 652-666., https://doi.org/10.3390/ai4030034.

[94] Usp. Rueda i dr., *»Just« accuracy? Procedural fairness…*, 6.

[95] Leilani H. GILPIN i dr., Explaining Explanations. An Overview of Interpretability of Machine Learning, u: Francesco Bonchi i dr. (ur.), *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, 80-89.

[96] Maria FOX, Derek LONG, Daniele MAGAZZENI, Explainable Planning, *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, London, King's College London, 2017, 24-30.

[97] Sherin M. MATHEWS, Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification. A Literature Review, u: Kohei Arai i dr. (ur.), *Advances in Intelligent*

»Osim toga, objašnjivost modela preduvjet je za izgradnju povjerenja i usvajanje sustava umjetne inteligencije.«[98]

Objašnjivost ima ključnu ulogu u definiranju odgovornosti za način na koji sustav UI-e funkcionira,[99] ali može biti i ključna u definiranju moralne odgovornosti koja je važna u situacijama kada nedostatak objašnjivosti može otežati definiciju moralne odgovornosti za donesene odluke.[100] Iz perspektive etičke provjere, autori navode da je objašnjivost dobar mehanizam upravljanja informiranjem pogođenih pojedinaca o tome kako je donesena odluka temeljena na predviđanju sustava UI.[101]

Do sada su se brojne metode OUI pokazale korisnim za ublažavanje i objašnjenje nepravednosti u situacijama predviđanja recidiva raka prostate i pomoću kojih su se uspješno razjasnili fenotipovi specifične za populacije afroameričkih pacijenata.[102] Takvi rezultati imaju višestruke pozitivne efekte na razvoj povjerenja jer objašnjenja pružaju koherentniju i djelotvorniju kliničku interpretaciju zdravstvenom djelatniku ili pacijentu.[103] Autori poput Fahmya i Csecsa se slažu da OUI ima značajan znanstveni potencijal jer može povećati povjerenje pacijenata, poboljšati komunikaciju i poboljšati sustave UI.[104] Ovaj pristup može demistificirati odluke UI, potičući bolje rezultate liječenja i poboljšavajući zajedničko donošenje odluka.[105] Također omogućuje personalizirane planove njege, značajno poboljšanje u odnosu na tradicionalne sustave UI.[106]

Autori Kiyaseeh, Laca i ostali prije spomenuti razvili su »training with explanations« (TWIX) kao dodatnu aplikaciju SAIS sustava, radi ublažavanje pristranosti i razvoja pravednih objašnjenja uspoređujući ih s objašnjenjima ljudskih stručnjaka. Iako je SAIS sustav pružao vizualno objašnjenje za svoju

Systems and Computing, Springer, 2019, 1269-1292.

[98] Krishna GADE i dr., Explainable AI in Industry, u: Ankur Teredesai i dr. (ur.), *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, Association for Computing Machinery, 2019, 3203-3204.

[99] Usp. Luciano FLORIDI i dr., AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, *Minds and machines*, 28 (2018) 4, 689-707, doi:10.1007/s11023-018-9482-5.

[100] Usp. Tsamodos i dr., *The ethics of algorithms...*, 19.

[101] Usp. Jakob MÖKANDER i dr., Conformity Assessments and Post-market Monitoring. A Guide to the Role of Auditing in the Proposed European AI Regulation, *Minds & Machines*, 32 (2022) 241-268, https://doi.org/10.1007/s11023-021-09577-4.

[102] Usp. James A. DIAO i dr., Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes, *Nature communications*, 12 (2021) 1, 1-15, doi:10.1038/s41467-021-21896-9.

[103] Usp. Chen, *Algorithmic fairness in artificial intelligence*…, 739.

[104] Usp. Ahmed S. FAHMY i dr., An Explainable Machine Learning Approach Reveals Prognostic Significance of Right Ventricular Dysfunction in Nonischemic Cardiomyopathy, *JACC. Cardiovascular imaging*, 15 (2022) 5, 766-779, doi:10.1016/j.jcmg.2021.11.029.

[105] Usp. Harshita PATEL i dr., Interactive XAI for personalized and trusted healthcare. Need of the hour, *International journal of surgery*, 110 (2024) 9, 5869-5870, doi:10.1097/JS9.0000000000001643.

[106] Usp. Giulia VILONE, Luca LONGO, Explainable Artificial Intelligence, *A Systematic Review*, 2020, ArXivabs/2006.00093.

procjenu vještina, objašnjavajući zašto je neka procjena vještina napravljena,[107] TWIX je dodatno unaprijeđen. TWIX može predvidjeti važnost video isječaka koji se koriste za procjenu kirurške vještine, rješavajući pristranost uzrokovanu neprovjerljivim video prikazima.[108] TWIX pruža kirurzima korisne povratne informacije pregledavanjem video zapisa kirurga i procjenom njihove razine vještina. Osim toga, TWIX u nekim slučajevima drastično ublažava pristranost objašnjenja koju pokazuje SAIS sustav. To je vidljivo iz testiranja poboljšanja putem testiranja »area under the precision recall curve« (AUPRC)[109] na temelju kojeg je TWIX pokazao bolje rezultate i unaprijedio pravednost putem objašnjenja. TWIX, kao poboljšana izvedba sustava UI, potencijal je za buduća poboljšanja sustava UI koju se predlažem za daljnje usavršavanje da bi se u budućnosti moglo razvijati efikasne strategije za ublažavanje pristranosti i razvoj objašnjivosti.

Zbog toga se smatra da je daljnji razvoj OUI ključan za razvoj povjerenja u primjenu sustava UI, osobito u zdravstvenom sektoru. Transparentnost i objašnjivost sustava omogućuju bolje razumijevanje mehanizama donošenja odluka, što je temelj za etičku i profesionalnu odgovornost. Ali OUI ne samo da pomaže u smanjenju pristranosti i nepravednosti u sustavima UI, već omogućuje zdravstvenim djelatnicima i pacijentima bolje razumijevanje i povjerenje u odluke koje UI donosi. Integracija objašnjivosti u dizajn sustava UI doprinosi pravednosti, omogućuje personaliziranu skrb i jača zajedničko donošenje odluka. Na taj način, OUI postaje ne samo tehnički već i etički alat koji osigurava bolje zdravstvene ishode i dugoročno povjerenje u tehnologiju.

Zaključak

Izazovi pristranosti i nepravednosti nastalih korištenjem sustava UI u zdravstvu mogu biti ublaženi metodama OUI. Sustavi UI mogu biti pristrani zbog povijesnih podataka ili dizajna modela, što može dovesti do nepravednih i netočnih predviđanja, posebno za društvene skupine određene rasnom, nacionalnom ili etničkom pripadnošću. U zdravstvu to može utjecati na dijagnosticiranje, liječenje i medicinske intervencije, a posljedice mogu biti životno ugrožavajuće. Pristranost može nastati iz više izvora, a ključ za rješenje ovih problema je u objašnjivosti i transparentnosti sustava UI. Objašnjivost omogućava zdravstvenim djelatnicima i pacijentima bolje razumijevanje odluka

[107] Usp. Dani KIYASSEH i dr., A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons, *Communications medicine*, 3 (2023) 1, 1-12, doi:10.1038/s43856-023-00263-3.

[108] Usp. Mittermaier i dr., *Bias in AI-based models for medical applications*…, 2.

[109] Usp. Jesse DAVIS, Mark GOADRICH, The relationship between precision-recall and ROC curves, u: William Cohen i dr. (ur.), *Proceedings of the 23rd International Conference on Machine learning*, New York, Association for Computing Machinery, 2006, 233-240.

donesenih od strane sustava UI, čime se povećava povjerenje i smanjuje rizik od donošenja pristranih i diskriminatornih odluka. Osim toga, OUI pomaže u razvoju povjerenja te pomaže u prepoznavanju i ublažavanju pristranosti.

Predlažući razvoj sustava UI kao što je TWIX, koji poboljšava povjerenje i ublažava pristranosti pružanjem objašnjenja, možemo pridonijeti razvoju sustava UI koji su pouzdani i etični. Primjena sustava poput TWIX-a pokazuje da tehnička rješenja mogu značajno smanjiti pristranost omogućavajući objašnjiva predviđanja, što pozitivno utječe na kliničke ishode i povjerenje u tehnologiju.

Daljnjim razvojem sustava OUI putem objašnjivih predviđanja omogućava se i pretpostavka informiranog pristanka, čime OUI nije samo tehničko rješenje, već i etički okvir za primjenu UI u zdravstvu, čineći je pravedenijom, pouzdanijom i učinkovitijom.

Luka Poslon*

Enhancing Trust by Ensuring Fairness in Medical AI

Summary

The paper addresses current issues of fairness in context of the application of artificial intelligence (henceforth: AI) in healthcare. Along with the difficulties in conceiving fairness, paper demonstrates the significance of explainable AI (henceforth: xAI), since xAI helps to mitigate bias and build trust in the use of Medical AI. Case studies of AI's use in healthcare serve as a reminder that an AI system may produce unfair and biased prediction results, which might exacerbate societal disparities based on race, ethnic, or other factors. Thus, AI systems that will mitigate bias and minimaze effects of discrimination—which can negatively impact on the decision-making process—must be developed. Even with technological solutions for reducing bias in AI systems, such as the *What-If Tool* and *AI Fairness 360*, additional effort is required to create AI that can provide explanations for its predictions in order to adhere to the ethical norms. Furthermore, explainability is especially important for healthcare in situations requiring informed consent. This was highlighted in the most recent Artificial Intelligence Act regulation, which placed a strong emphasis on explainability, transparency, and human oversight. This paper presents technical solutions that can significantly reduce bias by enabling explainable predictions, which positively impacts clinical outcomes and trust in the technology. Proposal for the development of AI systems like TWIX, which improve trust and mitigate bias by providing explanations, we can contribute to the development of future AI systems that are trustworthy and ethical. Therefore, I think that an AI system named TWIX, which guarantees the fairness of predictions due to its explainability capabilities, offers ways to reduce bias and discrimination. Future research should focus on developing similar AI systems with predictive explanatory abilities in order to promote fair and trustworthy decision-making.

Key words: artificial intelligence, explainability, fairness, healthcare, trust.

(*na engl. prev.* Luka Poslon)

* Luka Poslon, M.A., Catholic University of Croatia, Digital healthcare ethics laboratory (DigitHeaL); Address: Ilica 244, HR-10000, Zagreb, Croatia; e-mail: luka.poslon@unicath.hr.