UDK 004.8:614
614:004.8
https://doi.org/10.31337/oz.81.1.5
Pregledni rad
Primljeno: 28.3.2025.
Prihvaćeno: 19.11.2025.

Objašnjiva umjetna inteligencija i njezini izazovi kod donošenja odluka u zdravstvu

Luka Poslon[*]

Sažetak

U ovom radu autor usmjerava pozornost na potencijal i izazove primjene umjetne inteligencije u zdravstvu, s naglaskom na objašnjivu umjetnu inteligenciju, koja nudi transparentnost u donošenju odluka. Potencijal umjetne inteligencije u zdravstvu ostvaruje se u brojnim područjima, a sve se više raspravlja o etičkim izazovima poput objašnjivosti, pristranosti i povjerenja. Objašnjiva umjetna inteligencija može unaprijediti zdravstvenu skrb i ponuditi rješenje za specifične vrste pristranosti, što je naglašeno u ovom radu.

Ključne riječi: *umjetna inteligencija; objašnjiva umjetna inteligencija; zdravstvo; medicina; donošenje odluka; etika; pristranost*

Uvod

Velik napredak umjetne inteligencije (eng. *artificial intelligence*, AI) probudio je nadu kako će se potencijal AI moći ostvarivati u zdravstvu i biomedicinskim otkrićima za poboljšanje kvalitete zdravstvene skrbi. Višestruke koristi AI očituju se u zdravstvu u raznim područjima pri dijagnosticiranju bolesti, razvoju personaliziranih planova liječenja i podršci liječnicima pri donošenju odluka (Alowais et al., 2024). Iako AI igra ključnu ulogu u revoluciji zdravstva u 21. stoljeću, velik potencijal primjene sustava AI u zdravstvu i kliničkoj praksi još uvijek je ograničen. Prilikom donošenja odluka u zdravstvu, općenito se shvaća kako je pri opredjeljivanju za korištenje AI u zdravstvu potrebno preuzeti odgovornost za postupke onoga tko odlučuje, što zahtijeva evaluaciju ishoda sustava AI prije korištenja u kliničkoj praksi. U radu govorimo o suvremenim izazovima primjene AI u zdravstvu, kao i o granicama objašnjive umjetne inteligencije (eng. *explainable artificial intelligence*, XAI), kao nove metode u zdravstvu.

Studije slučaja podsjećaju nas koliko je važno propitivati predviđanja sustava AI (Gaube et al., 2021, 1). Reći kako imamo povjerenja u sustav AI više je nego

[*] Luka Poslon, mag. phil., doktorand, Hrvatsko katoličko sveučilište, Laboratorij za etiku digitalnih tehnologija u zdravstvu (Digit–HeaL). Adresa: Ilica 244, Zagreb, Hrvatska. ORCID iD: https://orcid.org/0000–0002–7389–7694. E–adresa: *luka.poslon@unicath.hr*

reći kako je taj sustav AI pouzdan. Pouzdanost je nedovoljna za stvaranje povjerenja, iako nam može posredovati razumijevanje rada sustava AI. Razumijevanje kako i zašto je sustav AI došao do određenoga predviđanja nije samo izazov na koji odgovaraju računalne znanosti, nego je za cjelovit odgovor potrebna interdisciplinarna suradnja. Potreba za objašnjivim i odgovornim predviđanjima povezana je s mogućnošću razumijevanja rada AI, a središnju ulogu u tumačenju razumijevanja imaju epistemologija i etika AI (Russo et al., 2023, 1585).

1. Etički izazovi umjetne inteligencije u zdravstvu

Brojna istraživanja pokazuju potencijal na temelju kojega sustavi AI mogu dostići ili premašiti učinak liječnika u specifičnim zadatcima (Liu et al., 2019, e294), kao što su točnost u dijagnozi upale srednjega uha na temelju otoskopije (Suresh et al., 2024, 1598), prepoznavanje abnormalnosti na rendgenskoj snimci prsnoga koša (Anderson et al., 2024) ili dijagnosticiranje i prognoza stadija karcinoma (Fountzilas et al., 2025, 9). Osim toga, potencijal se ogleda u pomoći pri izgradnju infrastrukture potrebne za brigu o sve starijoj populaciji, uporabi sve šireg znanja o bolestima i mogućnostima liječenja te borbi protiv nedostatka radne snage i iscrpljenosti medicinskih stručnjaka (Silcox et al., 2024, 1). Uz to, AI može unaprijediti dijagnostičku preciznost, optimizirati strategije liječenja te poboljšati skrb za pacijente putem personaliziranih intervencija i daljinskoga praćenja (Basubrin, 2025, 1). Usprkos velikomu potencijalu koji AI ima, primjena mora biti kontrolirana i komplementarna radu zdravstvenih djelatnika, jer ne može zamijeniti empatiju, holistički pristup i neposredan ljudski kontakt, što čini ključni dio ishoda liječenja prema nekim autorima (Montemayor et al., 2022, 1353). Osim toga, potencijal u zdravstvu trenutačno je teško ostvariti, kao što je prethodno navedeno, s obzirom na to da zdravstveni podatci potrebni za obuku, testiranje, korištenje i nadzor nisu standardizirani niti su uvijek dostupni. Osim toga, zdravstveni se podatci s vremenom mijenjaju, što dodatno otežava ostvarenje potencijala AI u zdravstvu (Silcox et al., 2024, 1). Eric Topol (2019, 17) upozorava kako se puno izazova u zdravstvu, koji se tiču društvenih i etičkih perspektiva, neće moći riješiti naprednom tehnologijom ili algoritmima.

U sklopu Uredbe 2024/1689 Europskoga parlamenta i Vijeća Europske unije od 13. lipnja 2024. o utvrđivanju usklađenih pravila o umjetnoj inteligenciji Stručna skupina za umjetnu inteligenciju razvila je sedam neobvezujućih etičkih načela za AI kojima se nastoji osigurati pouzdanost i etička prihvatljivost AI. Tih sedam načela uključuju ljudsko djelovanje i nadzor te transparentnost kako bi se osigurali etički standardi u skladu s vrijednostima Europske unije (EU, 2024).

Brojni sustavi AI koriste složene algoritme koji stvaraju predviđanja na temelju netransparentnih algoritamskih operacija koje većina liječnika teško može razumjeti i koristiti takva predviđanja u svojoj postojećoj praksu. Primjerice, neuronska mreža za klasifikaciju slika Inception v3 točnija je od liječnika u identificiranju dijabetičke retinopatije i raka kože jer uspijeva upravljati s malo manje od 25 milijuna parametara (Szegedy et al., 2016, 2818). Takva vrsta složenosti

znatno otežava razumijevanje načina rada sustava AI, zbog čega se taj fenomen opisuje kao "crna kutija". S obzirom na otežano razumijevanje i brojne praznine u razumijevanju rada netransparentnih sustava AI brojni znanstvenici upozoravaju na etičke i epistemološke izazove u kontekstu povjerenja u primjenu AI u zdravstvu (Wang et al., 2020, 59) kao pouzdanih medicinskih autoriteta (Wolkenstein, 2024, 371). Osim što liječnici teže mogu razviti povjerenje prema tehnologiji koju ne mogu objasniti, slični se izazovi očituju i u odnosu pacijent–liječnik, gdje brojni liječnici ne mogu pacijentu objasniti dijagnozu, prognozu ili terapiju za određenu bolest (Paranjape et al., 2019, 3). Mogućnost objašnjenja predviđanja sustava AI, s druge strane, omogućuje liječnicima prepoznavanje situacija u kojima se sustav AI primjenjuje za neprikladnu populaciju pacijenata. Također, objašnjenje predviđanja može pružiti upozorenje liječnicima ako su alati AI zastarjeli i potrebno ih je ažurirati (Giordano et al., 2021, 6). Pri pružanju odgovora na nove izazove koji se pojavljuju u kontekstu primjene AI u zdravstvu važno je poštivati moralne i pravne standarde. Sva buduća rješenja moraju biti ne samo etički odgovorna, nego i zakonski usklađena (Schneeberger et al., 2020, 9).

Suvremeni klinički sustavi AI za podršku pri donošenju odluka imaju ograničenja u svojoj primjeni na populaciju pacijenata, kontekstualne promjene i terapijske mogućnosti (Paranjape et al., 2019, 4). Zbog tih ograničenja budući će liječnici u svakodnevnoj praksi, osim razumijevanja načela medicine, također morati steći zadovoljavajuće znanje o matematičkim konceptima, osnovama AI te odgovarajućim etičkim i pravnim pitanjima. Te će im vještine pomoći u korištenju podatka iz širokoga spektra izvora, nadziranju alata AI i prepoznavanju slučajeva u kojima sustavi AI možda nisu točni kao što se to očekuje.

2. Utjecaj fenomena "crne kutije" na razvoj odgovornosti u zdravstvu

Brojni izazovi stvaraju prepreke za širu primjenu AI u zdravstvu, kao što su odobrenja regulatornih tijela, integracija sa sustavima elektroničkih zdravstvenih kartona, standardizacija kako bi slične tehnologije radile na sličan način, održavanje sustava AI i uređaja. Bohr i Memarzadeh (2020, 51) predviđaju kako će navedeni izazovi šire primjene sustava AI biti nadvladani do 2030 godine. No ostaje ključno pitanje, ako se navedeni izazovi uspiju nadvladati, može li primjena sustava AI koji funkcioniraju na načelu "crne kutije" u zdravstvu ugroziti sigurnost i kvalitetu liječničke skrbi za pacijente, posebice prilikom donošenja odluka? I podrazumijeva li kvalitetna liječnička skrb sposobnost objašnjenja medicinskih odluka?

2.1. Negativne posljedice fenomena "crne kutije" u zdravstvu

Specifičan fenomen "crne kutije" uzrok je izazovima koji su povezani s neprozirnošću i nedostatkom objašnjivosti te zaslužuju dodatnu pozornost kako bismo odgovorno rabili AI u zdravstvu. Osjetljivost pri donošenju odluka u zdravstvu naglašava imperativ za odgovornom uporabom AI u zdravstvu. Imperativ za od-

govornom primjenom AI u zdravstvu štiti povjerenje pacijenata, osigurava privatnost podataka i podržava načela medicinske etike (Upadhyay et al., 2023, 3). Većina sustava AI trenutačno rade po načelu "crne kutije", što onemogućava inženjerima, liječnicima ili pacijentima razumijevanje razloga zbog koji AI dolazi do određenih odluka ili predviđanja zbog netransparentnoga načina rada sustava AI. Sustavi AI rade po načelu "crne kutije" ako su njihovi unutarnji mehanizmi rada nerazumljivi čovjeku (Nicora et al., 2024, 1). O fenomenu "crne kutije" i nedostatku transparentnosti sustava AI raspravljalo se je u literaturi s obzirom na načine ublažavanja neprozirnosti s ciljem povećanja povjerenja u točnost AI (Grote, 2021, 337; Durán i Jongsma, 2021, 329). O neprozirnosti sustava AI također se je raspravljalo u kontekstu mogućnosti uvođenja medicinskoga paternalizma (McDougall, 2019, 157) te u kontekstu ugrožavanja niza važnih elemenata za pružanje skrbi, kao što su pravednost i autoritet liječnika (Grot i Berens, 2020, 207).

Do izazova povezanih s fenomenom "crne kutije" dolazi kada razlozi zbog kojih liječnik donosi odluku uz podršku AI nisu razumljivi pacijentu ili onima koji su uključeni u njegu pacijenta. Kao što smo spomenuli ranije, fenomen "crne kutije" je problematičan zbog otežanoga razumijevanja načela rada sustava AI, no "crna kutija" uzrokuje i izazove u razmjeni znanja i informacija između pacijenta i liječnika. Razmjena relevantnih informacija o pacijentovoj kliničkoj skrbi ključno je za razvoj povjerenja između pacijenta i liječnika. Proces donošenja odluka u zdravstvu pretpostavlja kako su razlozi za donošenje odluka zdravstvenomu osoblju razumljivi (Ali et al., 2023). Međutim, ono što ne razumijemo ne možemo adekvatno razmijeniti, zbog čega liječnik ne može opravdati svoje odluke temeljene na predviđanju sustava AI, što je potrebno za potpuno informiranje pacijenta o određenoj dijagnozi, terapiji ili prognozi bolesti. Osim toga, izazovi u tumačenju unutarnjih procesa sustava AI mogu biti opasni i dovesti do pristranih rezultata ili zaključaka koji ugrožavaju zdravlje pacijenta. Primjerice, IBM Watson For Oncology, klinički sustav AI za podršku pri donošenju odluka onkolozima, koji radi po načelu "crne kutije", kritiziran je zbog netočnih predviđanja i predviđanja štetnih zdravstvenih tretmana (Harish et al., 2021, 34).

2.2. XAI i odgovornost u zdravstvenoj skrbi

Odgovoran razvoj skrbi u sklopu navedenih izazova u zdravstvu treba uključivati i izazove i zabrinutosti koje pacijenti, kao korisnici, navode da su im važni. Kako bismo izgradili sustave AI kojima pacijenti vjeruju, potrebno je provesti standardizaciju alata AI u zdravstvu koji bi nam pomogli prilikom izračuna vjerojatnosti nastanka štete za pacijente uzrokovane odlukama koje donosi klinički alat temeljen na sustavu AI (Habli et al., 2020, 251). Prilikom standardizacije važno je voditi se kriterijima sigurnosti i odgovornosti, na čemu je potrebno raditi u budućnosti. Osiguravanje sigurnosti pacijenata jedan je od dva najčešća izazova o kojima se govori kada u zdravstvu u procesima donošenja odluka sudjeluje AI. S druge strane, razvoj i integracija AI mora biti popraćena adekvatnim razvojem moralne odgovornosti kako bismo osigurali da odluke donesene uz podršku

AI u budućnosti ne ugrožavaju zdravlje i živote pacijenata. Pitanje odgovornosti postaje ključno kada implementirana rješenja ugrožavaju zdravlje ili život pacijenata i dodatno produbljuju društvene nejednakosti ili uzrokuju nepravednost (Vela et al., 2022, 477).

Određivanje gdje je točno odgovornost u složenim društveno–tehničkim postupcima nije jednostavan zadatak i zajednički bi ga trebali rješavati svi dionici u procesu donošenja odluka. Trenutačne rasprave naglasak stavljaju na sigurnost pacijenta i ravnotežu odgovornosti između pojedinih zdravstvenih djelatnika i organizacija u kojima rade (Aveling et al., 2016, 217). S obzirom na to kako AI s razvojem postaje sve autonomnija, određivanje odgovornosti postaje još zahtjevnije, a tumačenje odgovornosti postaje dvosmisleno u situacijama kada krene po zlu. No, kao što tvrdi Virginia Dignum, istinski odgovorna AI ne može imati autonomiju bez nekoga oblika odgovornosti. Dignum je formirala ART načela (*Accountability, Responsibility, Autonomy*), kroz koja želi naglasiti važnost odgovornoga upravljanja sustavima AI koji se temelji na odgovornosti, transparentnosti i mogućnosti opravdanja odluka, i koji uključuje ljude, tehnologije i društvo u cjelokupni proces (Dignum, 2019, 53). Iz perspektive razvoja sustava AI, ART zahtijeva nove metode koje podržavaju integraciju etičkoga i društvenoga utjecaja sustava AI u proces inženjerstva.

Jedna od novih metoda, koja istovremeno poštuje etička načela, može unaprijediti zdravstvenu skrb na odgovoran način i pokušava riješiti izazove "crne kutije" pružanjem objašnjenja predviđanja sustava AI zove se objašnjiva umjetna inteligencija. XAI liječnicima može pružiti objašnjenja u obliku vizualizacija ili teksta, što omogućuje pouzdanije kliničke odluke umjesto da odluke temelje na automatiziranim predviđanjima (Amann et al., 2020). Osim toga, objašnjivost u sustavima AI u zdravstvu izravno doprinosi razvoju odgovorne uporabe s obzirom na to kako objašnjivost pruža sigurnosne provjere koje mogu biti korisne prilikom donošenja odluka (Upadhyay et al., 2023, 2).

Iako postoji nekoliko različitih definicija XAI, jedna od najčešćih označava XAI kao područje istraživanja koje se bavi objašnjenjem odluka sustava AI te ističe sustave koji su sposobni objasniti svoja predviđanja ili su poboljšani metodama objašnjenja (Finzel, 2025, 514). Specifično, cilj je XAI objasniti informacije koje dolaze od strane sustava koji rade po načelu "crne kutije" te objasniti način na koji netransparantan sustav donosi.

3. Donošenje odluka uz pomoć XAI

Kao što smo napomenuli, XAI je nastala kao rezultat potrage za odgovornim unaprjeđenjem zdravstvene skrbi. Cilj je XAI objasniti na transparentan način kako sustav AI donosi predviđanja, što pruža osnovu za razvoj povjerenja u primjenu AI u zdravstvu (Yang et al., 2022, 31). U zdravstvu, gdje se često donose visokorizične odluke sa životnim posljedicama, razumijevanje razloga iza odluka koje predlaže AI jedan je od ključnih elemenata za razvoj povjerenja u sustave AI. Međutim, sustavi AI koji rade po načelu "crne kutije" još se uvijek najčešće

primjenjuju u zdravstvu te zbog nemogućnosti objašnjenja predstavljaju problem za razvoj povjerenja.

3.1. Važnost povjerenja pri donošenju odluka

Unatoč činjenici da su robotski kirurški sustavi učinkoviti poput liječnika, mnogi pacijenti prilikom donošenja odluka još uvijek više vjeruju kirurgu nego robotskomu sustavu (Longoni et al., 2019, 641). Prema nekoliko autora, nedostatak objašnjivosti sustava AI uzrokuje smanjenje povjerenja u trenutačne dijagnostičke metode, kao što to može biti slučaj prilikom identifikacije karcinoma dojke (Bidwai et al., 2023, 150). S obzirom na sigurnost predviđanja odluka, brojni sustavi AI do sada nisu uspjeli uliti dovoljno povjerenja korisnicima sustava zdravstvene zaštite. S obzirom na to kako zdravstveni podatci mogu biti vrlo iskrivljeni i šumoviti, objašnjivost je jedna od najvažnijih komponenti za izgradnju povjerenja korisnika. Uz to, nedavno istraživanje naglasilo je sumnje zdravstvenih djelatnika u primjenu AI na jedinicama intenzivnoga liječenja. U istraživanju je 71% sudionika bilo nesigurno ili se nije slagalo s činjenicom kako se AI koja radi po načelu "crne kutije" može pouzdano koristiti u donošenju odluka u intenzivnom liječenju (Sande et al., 2022, 1815). Izraziti nedostatak povjerenja mogao bi biti uzrokovan nepovjerenjem koje zdravstveno osoblje ima prema predviđanjima temeljenim na sustavima AI koji nalikuju "crnim kutijama" (Abgrall et al., 2024). To otvara sljedeća pitanja: Trebamo li sustave AI učiniti objašnjivima liječnicima i može li XAI unaprijediti zdravstvenu skrb?

3.2. Unaprjeđuje li XAI zdravstvenu skrb?

Holzinger i suradnici ističu nužnu potrebu zdravstvenih djelatnika da prilikom donošenja odluka imaju mogućnost razumjeti kako i zašto je donesena određeno predviđanje sustava AI. Ta potreba dolazi do izražaja u zdravstvenom kontekstu koji je često neizvjestan i susreće se s vjerojatnostima te s nepoznatim, nepotpunim, neuravnoteženim, heterogenim, pogrešnim, netočnim i nedostajućim skupovima podataka (Holzinger et al., 2019, 2). Sustavi XAI mogu unaprijediti procese donošenja odluka u zdravstvu time što objašnjenja pomažu u predviđanju rizika od bolesti (Niu et al., 2024, 10). Osim toga, sustavi XAI mogu unaprijediti zdravstvenu skrb omogućavajući rana upozorenja za životno ugrožavajuća stanja, kao što su sepsa, akutna ozljeda bubrega ili akutna ozljeda pluća (Lauritsen et al., 2020, 2). Sustav XAI koji su razvili Lauritsen i suradnici pod nazivom *explainable AI early warning score*, ili skraćeno XAI–EWS, pruža vizualna objašnjenja za dana predviđanja u stvarnom vremenu te omogućuje uvid u elektroničke zdravstvene zapise koji inače ne bi bili identificirani (Lauritsen et al., 2020, 5). Takav pristup omogućuje liječnicima učinkovitiji i odgovorniji pristup svakomu pacijentu te unaprjeđuje zdravstvenu skrb.

Jedan od ključnih izazova koji se pojavljuje prilikom donošenja odluka u zdravstvu uz podršku sustava AI tiče se pristranosti. Sustavi AI mogu svojim predviđanjima primjerice na temelju pristranosti povećavati postojeće nejednakosti u

socioekonomskom statusu, rasi, etničkoj pripadnosti, vjeri, spolu ili invaliditetu, što je dobro poznato u literaturi (Mittermaier et al., 2023, 1). Primjeri pristranosti zabilježeni su u brojnim situacijama, a mogu se dogoditi primjerice prilikom procjene tehničkih vještina liječnika (Kiyasseh et al., 2023a, 7) ili navođenja kirurga putem računalnoga vida tijekom operacije (Kiyasseh et al., 2023b, 789). Upravo su metode XAI korisne za prepoznavanje pristranosti u sustavima AI. Budući da se XAI oslanja na podatke za obuku, može prepoznati prekomjernu zastupljenost određene demografske skupine koja nije reprezentativna za ciljanu populaciju, što se naziva pristranost uzorkovanja (Upadhyay et al., 2023, 2). Dokazano je i da XAI može pomoći prilikom otkrivanja i ublažavanja pristranosti putem metode XAI pod nazivom "objašnjenje primjerom" (eng. *explanation by example*). Ta metoda temeljem efikasnoga prepoznavanja pristranosti omogućuje zaštitu od donošenja pogrešnih odluka (Hooker, 2021, 1).

Navedeni primjeri potvrđuju kako XAI može predviđanja učiniti razumljivijima za ljudske korisnike i pomoći prilikom donošenja odluka. To u konačnici može rezultirati poboljšanjem terapeutskoga odnosa i potaknuti povjerenje u primjenu AI u zdravstvu. No, osim toga, postoje različite granice XAI koje su kritičari u dosadašnjem radu istaknuli i kojima je potrebno posvetiti pažnju.

4. Ograničenja XAI u zdravstvu

Jedno od najčešćih ograničenja XAI tiče se kompromisa između točnosti i objašnjivosti predviđanja. Rasprostranjeno je uvjerenje kako složeniji sustavi AI koji rade po načelu "crne kutije" pružaju točnija predviđanja, a sustavi XAI zbog pružanja objašnjenja moraju raditi kompromise te pružaju predviđanja manje točnosti. Međutim, Rudin tvrdi da to često nije točno, osobito kada sustav XAI može obrađivati dobro strukturirane podatke. Čak i kada uspoređujemo sustave AI za računalni vid, gdje je dubinsko učenje vrlo učinkovito, a teško je definirati objašnjivost, moguće je postići objašnjivost bez gubitka točnosti (Rudin, 2019, 209). Istraživanje je pokazalo da javnost prilikom procjene medicinskih scenarija prednost daje točnosti zbog potrebe za točnim i pravovremenim odlukama za bolje ishode. Suprotno tomu, ispitanici su u scenarijima vezanim uz kazneno pravosuđe prednost dali objašnjivosti zbog osiguravanja pravednosti (Veer et al., 2021, 2137). Razlika u prioritetima između objašnjivosti i pravednosti u različitim društvenim sektorima naglašava potrebu za daljnjim razvojem smjernica i politike AI.

Drugo često ograničenje XAI tiče se nedostatka mjerljivih alata za procjenu kvalitete objašnjivosti. Iako su pojedina zakonska rješenja, kao što je Zakon o umjetnoj inteligenciji, pokrenuli proces standardizacije mjerenja objašnjivosti (EU, 2024), još uvijek nisu dovoljno jasne obveze vezane za objašnjivost te tko ih mora ispuniti (Sovrano et al., 2022, 136). Valja imati na umu kako predetaljna objašnjenja mogu biti preizazovna za razumijevanje, dok pojednostavljena objašnjenja mogu izgubiti točnost.

Treće ograničenje XAI tiče se istraživanja utjecaja pristranosti i objašnjenja u procesima automatizacija. U literaturi se većinom spominje jedna vrsta pristranosti koja se pojavljuje na temelju automatizacije kada se korisnik pretjerano oslanja na sustav AI (Nguyen, 2024, 1). Iako je potrebno dodatno istražiti korelaciju između pristranosti automatizacije i objašnjivosti, poseban slučaj pristranosti relevantan za zdravstvo tiče se pristranosti temeljenoj na prekomjernom samopouzdanju (Swofford i Champod, 2021, 4). Pristranost prekomjernoga samopouzdanja povezana je s konceptom algoritamske odbojnosti koji se javlja posebno kod stručnjaka koji odbijaju vjerovati predviđanjima sustava AI u određenom području, kao što se to može dogoditi na primjeru pilota. Zbog algoritamske odbojnosti, pojedinci svoje prosudbe tretiraju kao superiorne u odnosu na tuđe te razvijaju pristranost prekomjernoga samopouzdanja. Arkes i suradnici koji su proveli istraživanje prije gotovo 40 godina istaknuli su da stručnjaci zbog prekomjernoga samopouzdanja češće ignoriraju sustave AI, postižu lošije rezultate, vjeruju da su u pravu, ali na kraju štete onima kojima žele pomoći (Arkes et al., 1986, 107). Novija literatura sugerira da pretjerano povjerenje liječnika u preporuke sustava AI, posebno kada one nisu objašnjene na razumljiv način, može negativno utjecati na kliničko odlučivanje, što može povećati rizik od medicinskih pogrešaka (Al–Maghrabi et al., 2024, 133). Također, novija istraživanja navode kako je trećina pružatelja zdravstvenih usluga u ruralnom Senegalu previše samouvjerena te pretjerano samopouzdani pružatelji usluga imaju 26% manju vjerojatnost da će pružati ispravnu skrb svojim pacijentima (Kovacs et al., 2020, 3).

Metode XAI već su dokazano pomogle pri prepoznavanju i ublažavanju pristranosti te stoga predlažem razvoj metoda XAI koje će biti usmjerene na pristranosti prekomjernoga samopouzdanja liječnika i zdravstvenoga osoblja. Takve metode XAI trebaju imati uključene mjere nesigurnosti koje korisnike potiču na opreznije i nijansiranije tumačenje rezultata sustava AI, priznavajući ograničenja objašnjenja (Payrovnaziri et al., 2020, 1176). Jedna od ključnih prednosti pristupa koji uključuje metode XAI i mjere nesigurnosti sposobnost je ublažavanja potencijalnih pristranosti u korisničkom tumačenju. Iako se takav pristup još nije koristio, potencijalno može potaknuti dublje povjerenje u rezultate sustava AI i olakšati usvajanje sličnih metoda u zdravstvenim aplikacijama. Osim toga, posljedice pristranosti prekomjernoga samopouzdanja u području zdravstva mogle bi biti smanjenje ako putem objašnjivosti dobijemo važne uvide koje pomažu upravo stručnjacima za unaprjeđenje zdravstvene skrbi.

Cilj primjene AI u zdravstvu je pomoći svim korisnicima u procjeni jesu li predviđanja AI istinita, što je osobito važno u situacijama visokoga rizika. Uz to, većina sustava AI namijenjena je stručnjacima koji trebaju tumačiti određene informacije ili predviđanja, što ulogu liječnika čini još važnijom. Za pretpostaviti je kako će objašnjivost bitno pomoći u prepoznavanju i uklanjanju pristranosti prekomjernoga samopouzdanja, što je cilj svih dionika u zdravstvenoj skrbi, posebice pacijenata. Zbog toga je potrebno uložiti dodatne napore kako bi se istražile pristranosti prekomjernoga samopouzdanja i unaprijedila zdravstvena skrb.

Zaključak

Zaključno, na temelju analize literature i razmatranja osobnih stavova može se izdvojiti da, usprkos velikomu potencijalu AI u zdravstvu, primjena se suočava s velikim izazovima, osobito u pogledu odgovornosti i objašnjivosti predviđanja. Fenomen "crne kutije" jedan je od ključnih izazove, jer otežava povjerenje i komunikaciju između liječnika, pacijenta i sustava AI. Objašnjenje predviđanja od strane sustava AI ključno je za razvoj povjerenja, što omogućuje odgovorniji pristup i napredak zdravstvene skrbi. XAI ima mogućnosti ponuditi konkretne alate za rješavanje navedenih izazova, omogućujući veću transparentnost sustava, identifikaciju pristranosti i smanjenje rizika povezanih s prekomjernim samopouzdanjem algoritama. Na taj način, razvoj XAI predstavlja velik korak naprijed u prevladavanju tih prepreka. XAI može unaprijediti zdravstvenu skrb i doprinijeti odgovornoj primjeni AI u zdravstvu.

Usprkos granicama XAI važno je kontinuirano raditi na standardizaciji alata, usklađivanju s moralnim i zakonskim standardima te edukaciji zdravstvenih djelatnika u pogledu novih tehnologija. Povjerenje i odgovornost ključni su čimbenici za integraciju XAI u svakodnevnu praksu.

Samo odgovoran razvoj i primjena tehnologija, koje uzimaju u obzir ljudski faktor i specifičnosti svakoga pacijenta, mogu osigurati sigurno i učinkovito korištenje AI u zdravstvu. XAI predstavlja jedan od ključnih elemenata toga razvoja, omogućujući sustavima AI da budu ne samo tehnološki napredni, nego i etički i društveno odgovorni.

Literatura

Abgrall, Guillaume; Holder, Alexander L.; Chelly Dagdia, Zohra; Zeitouni, Karim; Monnet, Xavier (2024). Should AI models be explainable to clinicians? *Critical Care*, 28, no. 301, https://doi.org/10.1186/s13054–024–05005–y.

Ali, Sajid; Abuhmed, Tamer; El–Sappagh, Shaker; Muhammad, Khan; Alonso–Moral, José M.; Confalonieri, Roberto; Guidotti, Riccardo; Del Ser, Javier; Díaz–Rodríguez, Natalia; Herrera, Francisco (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, no. 101805, https://doi.org/10.1016/j.inffus.2023.101805.

Al–Maghrabi, Mohsin; Mamede, Silvia; Schmidt, Henk G.; Omair, Aamir; Al–Nasser, Sami; Alharbi, Nouf Sulaiman; Magzoub, Mohi Eldin Mohammed Ali (2024). Overconfidence, time–on–task, and medical errors: Is there a relationship? *Advances in Medical Education and Practice*, 15, 133–140, https://doi.org/10.2147/AMEP.S442689.

Alowais, Shuroug A.; Alghamdi, Sahar S.; Alsuhebany, Nada; Alqahtani, Tariq; Alshaya, Abdulrahman I.; Almohareb, Sumaya N.; Aldairem, Atheer; Alrashed, Mohammed; Bin Saleh, Khalid; Badreldin, Hisham A.; Al Yami, Majed S.; Al Harbi, Shmeylan; Albekairy, Abdulkareem M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, 23, no. 689, https://doi.org/10.1186/s12909–023–04698–z.

Amann, Julia; Blasimme, Alessandro; Vayena, Effy; Frey, Dietmar; Madai, Vince I.; Precise4Q Consortium (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, no. 310, https://doi.org/10.1186/s12911–020–01332–6.

Anderson, Pamela G.; Tarder–Stoll, Hannah; Alpaslan, Mehmet; Keathley, Nora; Levin, David L.; Venkatesh, Srivas; Bartel, Elliot; Sicular, Serge; Howell, Scott; Lindsey, Robert V.; Jones, Rebecca M. (2024). Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X–rays. *Scientific Reports*, 14, no. 25151, https://doi.org/10.1038/s41598–024–76608–2.

Arkes, Hershey R.; Dawes, Robyn M.; Christensen, Charles (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37(1), 93–110, https://doi.org/10.1016/0749–5978(86)90046–4.

Aveling, Elizabeth L.; Parker, Mary; Dixon-Woods, Mira (2016). What is the role of individual accountability in patient safety?: A multi–site ethnographic study. *Sociology of Health & Illness*, 38(2), 216–232, https://doi.org/10.1111/1467–9566.12370.

Basubrin, Omar (2025). Current status and future of artificial intelligence in medicine. *Cureus*, 17(1), no. e77561, https://doi.org/10.7759/cureus.77561.

Bidwai, Pooja; Khairnar, Smita; Gite, Shilpa (2023). Explainable artificial intelligence in breast cancer identification. U: *Chapman and Hall/CRC eBooks* (str. 148–165). https://doi.org/10.1201/9781003333425–8.

Bohr, Adam; Memarzadeh, Kaveh (2020). The rise of artificial intelligence in healthcare applications.U: A. Bohr i K. Memarzadeh(ur.), *Artificial Intelligence in Healthcare* (str. 25–60). Academic Press.

Dignum, Virginia (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer.

Durán, Juan Manuel; Jongsma, Karin Rolanda (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, 47(5), 329–335, https://doi.org/10.1136/medethics–2020–106820.

EU (2024). Uredba (EU) 2024/1689 Europskog parlamenta i Vijeća od 13. lipnja 2024. o utvrđivanju usklađenih pravila o umjetnoj inteligenciji. *EUR–Lex*, https://eur–lex.europa.eu/legal–content/HR/TXT/?uri=CELEX%3A32024R1689.

Finzel, Bettina (2025). Current methods in explainable artificial intelligence and future prospects for integrative physiology. *Pflugers Archiv*, 477(4), 513–529, https://doi.org/10.1007/s00424–025–03067–7.

Fountzilas, Elena; Pearce, Tillman; Baysal, Mehmet A.; Chakraborty, Abhijit; Tsimberidou, Apostolia M. (2025). Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. *NPJ Digital Medicine*, 8, no. 75, https://doi.org/10.1038/s41746–025–01471–y.

Gaube, Simon; Suresh, Harini; Raue, Moritz; Merritt, Andrew; Berkowitz, Stephen J.; Lermer, Eva; Coughlin, Jeffrey F.; Guttag, John V.; Colak, Emre; Ghassemi, Marzyeh (2021). Do as AI say: susceptibility in deployment of clinical decision–aids. *NPJ Digital Medicine*, 4, no. 31, https://doi.org/10.1038/s41746–021–00385–9.

Giordano, Cristiano; Brennan, Michael; Mohamed, Bassel; Rashidi, Peyman; Modave, Florent; Tighe, Patrick (2021). Accessing artificial intelligence for clinical decision–making. *Frontiers in Digital Health*, 3, no. 645232, https://doi.org/10.3389/fdgth.2021.645232.

Grote, Thomas (2021). Trustworthy medical AI systems need to know when they don't know. *Journal of Medical Ethics*, 47(5), 337–338, https://doi.org/10.1136/medethics–2021–107463.

Grote, Thomas; Berens, Philipp (2020). On the ethics of algorithmic decision–making in healthcare. *Journal of Medical Ethics*, 46(3), 205–211, https://doi.org/10.1136/medethics–2019–105586.

Habli, I Ibrahim; Lawton, Tim; Porter, Zoe (2020). Artificial intelligence in health care: Accountability and safety. *Bulletin of the World Health Organization*, 98(4), 251–256, https://doi.org/10.2471/BLT.19.237487.

Harish, Vinyas; Morgado, Felipe; Stern, Ariel D.; Das, Sunit (2021). Artificial intelligence and clinical decision making: The new nature of medical uncertainty. *Academic Medicine*, 96(1), 31–36, https://doi.org/10.1097/ACM.0000000000003707.

Holzinger, Andreas; Langs, Georg; Denk, Helmut; Zatloukal, Kurt; Müller, Heimo (2019). Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews Data Mining and Knowledge Discovery*, 9(4), no. e1312, https://doi.org/10.1002/widm.1312.

Hooker, Sara (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4), no. 100241, https://doi.org/10.1016/j.patter.2021.100241.

Kiyasseh, Dani; Laca, Jasper; Haque, Taseen F.; Miles, Brian J.; Wagner, Christian; Donoho, Daniel A.; Anandkumar, Animashree; Hung, Andrew J. (2023a). A multi–institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. Communications Medicine, 3(1), no. 42, https://doi.org/10.1038/s43856–023–00263–3.

Kiyasseh, Dani; Ma, Rui; Haque, Taseen F.; Miles, Brian J.; Wagner, Christian; Donoho, Daniel A.; Anandkumar, Animashree; Hung, Andrew J. (2023b). A vision transformer for decoding surgeon activity from surgical videos. *Nature Biomedical Engineering*, 7(6), 780–796, https://doi.org/10.1038/s41551–023–01010–8.

Kovacs, Roxanne J.; Lagarde, Mylène; Cairns, John A. (2020). Overconfident health workers provide lower quality healthcare. *Journal of Economic Psychology*, 76, no. 102213, https://doi.org/10.1016/j.joep.2019.102213.

Lauritsen, Simon Meyer; Kristensen, Mads; Olsen, Mathias Vassard; Larsen, Morten Skaarup; Lauritsen, Katrine Meyer; Jørgensen, Marianne Johansson; Lange, Jeppe; Thiesson, Bo (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, 11(1), no. 3852, https://doi.org/10.1038/s41467–020–17431–x.

Liu, Xiaoxuan; Faes, Livia; Kale, Aditya U.; Wagner, Siegfried K.; Fu, Dun Jack; Bruynseels, Alice; Mahendiran, Thushika; Moraes, Gabriella; Shamdas, Mohith; Kern, Christoph; Ledsam, Joseph R.; Schmid, Martin K.; Balaskas, Konstantinos; Topol, Eric J.; Bachmann, Lucas M.; Keane, Pearse A.; Denniston, Alastair K. (2019). A comparison of deep learning performance against health–care professionals in detecting diseases from medical imaging: a systematic review and meta–analysis. *The Lancet Digital Health*, 1(6), e271–e297, https://doi.org/10.1016/S2589–7500(19)30123–2.

Longoni, Chiara; Bonezzi, Andrea; Morewedge, Carey K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4), 629–650, https://doi.org/10.1093/jcr/ucz013.

McDougall, Rosalind J. (2019). Computer knows best? The need for value–flexibility in medical AI. *Journal of Medical Ethics*, 45(3), 156–160, https://doi.org/10.1136/medethics–2018–105118.

Mittermaier, Mirja; Raza, Marium M.; Kvedar, Joseph C. (2023). Bias in AI–based models for medical applications: Challenges and mitigation strategies. *NPJ Digital Medicine*, 6, no. 113, https://doi.org/10.1038/s41746–023–00858–z.

Montemayor, Carlos; Halpern, Jodi; Fairweather, Abrol (2022). In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI and Society*, 37(4), 1353–1359, https://doi.org/10.1007/s00146–021–01230–z.

Nguyen, Tina (2024). ChatGPT in Medical Education: A Precursor for Automation Bias? *JMIR Medical Education*, 10, no. e50174, https://doi.org/10.2196/50174.

Nicora, Giovanna; Catalano, Michele; Bortolotto, Chandra; Achilli, Marina Francesca; Messana, Gaia; Lo Tito, Antonio; Consonni, Alessio; Cutti, Sara; Comotto, Federico; Stella, Giulia Maria; Corsico, Angelo; Perlini, Stefano; Bellazzi, Riccardo; Bruno, Raffaele; Preda, Lorenzo (2024). Bayesian networks in the management of hospital admissions: A comparison between explainable AI and black box AI during the pandemic. *Journal of Imaging*, 10(5), no. 117, https://doi.org/10.3390/jimaging10050117.

Niu, Shuai; Yin, Qing; Ma, Jing; Song, Yunya; Xu, Yida; Bai, Liang; Pan, Wei; Yang, Xian (2024). Enhancing healthcare decision support through explainable AI models for risk prediction. *Decision Support Systems*, 181, no. 114228, https://doi.org/10.1016/j.dss.2024.114228.

Paranjape, Ketan; Schinkel, Michiel; Nannan Panday, Rishi S.; Car, Josip; Nanayakkara, Prabath W. B. (2019). Introducing artificial intelligence training in medical education. *JMIR Medical Education*, 5(2), no. e16048, https://doi.org/10.2196/16048.

Payrovnaziri, Sahar N.; Chen, Zhen; Rengifo-Moreno, Pilar; Miller, Terry; Bian, Jiajie; Chen, J. H.; Liu, Xiaoqian; He, Zheng (2020). Explainable artificial intelligence models using real–world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7), 1173–1185, https://doi.org/10.1093/jamia/ocaa053.

Rudin, Cynthia (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215, https://doi.org/10.1038/s42256–019–0048–x.

Russo, Federica; Schliesser, Eric; Wagemans, Jean (2023). Connecting ethics and epistemology of AI. *AI and Society*, 39(4), 1585–1603, https://doi.org/10.1007/s00146–022–01617–6.

Sande, Davy van de; Genderen, Michel E. van; Braaf, H.; Gommers, D.; Bommel, Jasper van (2022). Moving towards clinical use of artificial intelligence in intensive care medicine: business as usual? *Intensive Care Medicine*, 48(12), 1815–1817. https://doi.org/10.1007/s00134–022–06910–y.

Schneeberger, David; Stöger, Karl; Holzinger, Andreas (2020). The European Legal Framework for Medical AI. U A. Holzinger et al. (ur.), M*achine Learning and Knowledge Extraction: Lecture Notes in Computer Science 12279*. Cham: Springer, https://doi.org/10.1007/978–3–030–57321–8_12.

Silcox, Christina; Zimlichmann, Eyal; Huber, Katie; Rowen, Neil; Saunders, Robert; McClellan, Mark; Kahn, Charles N. (III); Salzberg, Claudia A.; Bates, David W. (2024). The potential for artificial intelligence to transform healthcare: perspectives from international health leaders. *NPJ Digital Medicine*, 7, no. 88, https://doi.org/10.1038/s41746–024–01097–6.

Sovrano, Francesco; Sapienza, Salvatore; Palmirani, Monica; Vitali, Fabio (2022). Metrics, explainability and the European AI act proposal. *J*, 5(1), 126–138, https://doi.org/10.3390/j5010010.

Suresh, Krish; Wu, Michael P.; Benboujja, Fouzi; Christakis, Barbara; Newton, Alice; Hartnick, Christopher J.; Cohen, Michael S. (2024). AI model versus clinician otoscopy in the operative setting for otitis media diagnosis. *Otolaryngology — Head and Neck Surgery*, 170(6), 1598–1601, https://doi.org/10.1002/ohn.559.

Swofford, Henry; Champod, Christophe (2021). Implementation of algorithms in pattern & impression evidence: A responsible and practical roadmap. Forensic science international. *Synergy*, 3, no. 100142, https://doi.org/10.1016/j.fsisyn.2021.100142.

Szegedy, Christian; Vanhoucke, Vincent; Ioffe, Sergey; Shlens, Jonathon; Wojna, Zbigniew (2015). Rethinking the Inception Architecture for Computer Vision. U: *2016*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (str. 2818–2826). https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.308.

Topol, Eric J. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

Upadhyay, Umashankar; Gradisek, Anton; Iqbal, Usman; Dhar, Eshita; Li, Yu-Chuan; Syed-Abdul, Shabbir (2023). Call for the responsible artificial intelligence in the healthcare. *BMJ Health & Care Informatics*, 30, no. e100920, https://doi.org/10.1136/bmjhci–2023–100920.

Veer, Sabine N. van der; Riste, Lisa; Cheraghi-Sohi, Sudeh; Phipps, David L.; Tully, M. P.; Bozentko, K.; Atwood, S.; Hubbard, A.; Wiper, C.; Oswald, M.; Peek, Niels (2021). Trading off accuracy and explainability in AI decision–making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*, 28(10), 2128–2138, https://doi.org/10.1093/jamia/ocab127.

Vela, Monica B.; Erondu, Amarachi I.; Smith, Nichole A.; Peek, Monica E.; Woodruff, James N.; Chin, Marshall H. (2022). Eliminating explicit and implicit biases in health care: Evidence and research needs. *Annual Review of Public Health*, 43, 477–501, https://doi.org/10.1146/annurev–publhealth–052620–103528.

Wang, Fei; Kaushal, Rainu; Khullar, Dhruv (2020). Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Annals of Internal Medicine*, 172, 59–60, https://doi.org/10.7326/M19–2548.

Wolkenstein, Andreas (2024). Healthy mistrust: Medical black box algorithms, epistemic authority, and preemptionism. *Cambridge Quarterly of Healthcare Ethics*, 33(3), 370–379, doi:10.1017/S0963180123000646.

Yang, Guang; Ye, Qinghao; Xia, Jun (2022). Unbox the black–box for the medical explainable AI via multi–modal and multi–centre data fusion: A mini–review, two showcases and beyond. An *International Journal on Information Fusion*, 77, 29–52, https://doi.org/10.1016/j.inffus.2021.07.016.

Challenges of Explainable Artificial Intelligence in Healthcare Decision-Making Processes

Luka Poslon[*]

Summary

In this paper the author draws our attention to the potential and the challenges of the application of artificial intelligence in healthcare. Artificial intelligence offers advancements such as better diagnostics and personalized treatments, but also raises ethical challenges related to explainability, bias, and trust. A key challenge is the "black box" phenomenon, where artificial intelligence predictions lack transparency and cannot be easily explained. To address this, explainable artificial intelligence aims to make predictions explainable and foster trust in use of artificial intelligence in healthcare. However, this approach faces its own hurdles, such as balancing between accuracy and explainability, the lack of standardized tools for measuring ex-

[*] Luka Poslon, mag. phil., Ph.D. Student, Digital Healthcare Ethics Laboratory (Digit–HeaL), Catholic University of Croatia. Address: Ilica 244, Zagreb, Croatia. E–mail: *luka.poslon@unicath.hr*

planation quality, and under–researched relationship between bias and explanations in automated systems. Understanding these limitations is vital for the responsible use of artificial intelligence in high–risk healthcare settings. Special attention is given to the overconfidence bias, where artificial intelligence systems or their users may overestimate the reliability of outputs. The paper proposes an original explainable artificial intelligence method designed to detect and mitigate the effects of overconfidence bias, thereby reducing decision–making errors. By identifying such risks and addressing them through tailored an explainable artificial intelligence approach, the paper contributes to the development of trustworthy artificial intelligence systems in healthcare.

Keywords: *artificial intelligence; explainable artificial intelligence; healthcare; medicine; decision–making; ethics; bias*