Umjetna inteligencija, medicina i autonomija

Tomislav Bracanović*
tbracanovic@ifzg.hr
https://orcid.org/0000-0001-8168-2194

https://doi.org/10.31192/np.19.1.5
UDK: 004.8:[61:57.08]
17.025.2:61
Pregledni članak / Review
Primljeno: 8. listopada 2020.
Prihvaćeno: 25. studenog 2020.

U radu se razmatra pitanje jesu li na umjetnoj inteligenciji utemeljeni dijagnostički sustavi (AIBDS) prijetnja autonomiji pacijenata kao jednom od središnjih načela biomedicinske etike. U prvom dijelu rada objašnjava se što su AIBDS i kako funkcioniraju, s naglaskom na tehnologiji strojnog učenja i na njegovu svojstvu netransparentnosti. U drugom dijelu rada prikazuje se više paradigmatskih stavova o AIBDS-u kao prijetnji autonomiji pacijenata, uz osvrt na standardnu analizu prema kojoj ta autonomija uključuje (minimalno) tri komponente: intencionalnost, slobodu od ograničenjâ i razumijevanje. U trećem se dijelu rada argumente pro et contra teze o AIBDS-u kao prijetnji autonomiji razmatra iz perspektive pacijenata, dok ih se u četvrtom dijelu rada razmatra iz perspektive liječnika. U radu se pokazuje koliko rasprava o ovom pitanju može biti složena i koja oruđa argumentacije suprotstavljenim stranama pritom mogu stajati na raspolaganju.

Ključne riječi: *autonomija, biomedicinska etika, dijagnostički sustavi, strojno učenje, netransparentnost, umjetna inteligencija.*

Uvod

Medicina se sve više oslanja na »pragmatičnu« ili »inženjerski« orijentiranu umjetnu inteligenciju s njezinim ograncima kao što su procesiranje prirodnog jezika, reprezentacija znanja, automatizirano zaključivanje, strojno učenje, računalni vid i robotika.[1] Jedna od najpoznatijih primjena umjetne inteligencije

* Dr. sc. Tomislav Bracanović, znanstveni savjetnik, Institut za filozofiju, Ulica grada Vukovara 54, Zagreb.

[1] Usp. Stuart J. RUSSELL, Peter NORVIG, *Artificial Intelligence. A Modern Approach*, Harlow, Pearson, 2016, 2-3.

u medicini je korištenje robota za minimalno invazivne kirurške zahvate, prije svega radi njihove preciznosti koja uvelike nadilazi preciznost i najvještijih ljudi kirurga. No, umjetna inteligencija prožima medicinu u mnogo širem smislu – samo neki od primjera su razni »pametni« (*smart*) odjevni predmeti i dodaci (npr. satovi, narukvice i pojasevi koji prate razne tjelesne funkcije), različite vrste implantata, umjetnih udova i organa, tablete koje »znaju« kada trebaju početi djelovati, roboti za rehabilitaciju i rad s pacijentima s tjelesnim ili mentalnim oboljenjima, dronovi za dostavu doza krvi i sl.[2]

Medicinske primjene umjetne inteligencije, dakako, otvaraju i brojna etička pitanja, poput sigurnosti, pouzdanosti, odgovornosti za štete, neovlaštene upotrebe osobnih podataka, narušavanja privatnosti, pravedne distribucije dobrobitî koje donose i sl.[3] Potrebno je uočiti, međutim, da su ova etička pitanja uvelike »generička«, odnosno da ne prožimaju samo primjenu umjetne inteligencije u medicini nego i njezinu primjenu u mnogim drugim područjima ljudskog života, poput prometa, internetske trgovine, proizvodnje, upravljanja kućanstvima (*smart homes*) i gradskom infrastrukturom (*smart cities*) i sl. Ovaj se rad pak usredotočuje na pitanje za koje se može reći da je specifično upravo za medicinsku primjenu umjetne inteligencije. To pitanje glasi: Jesu li na umjetnoj inteligenciji utemeljeni dijagnostički sustavi (*artificial intelligence-based diagnostic systems*, u daljnjem tekstu: AIBDS)[4] prijetnja autonomiji pacijenata?

1. Što je AIBDS?

AIBDS su ekspertni sustavi namijenjeni otkrivanju različitih vrsta oboljenja na temelju velikih skupova podataka koji u suvremenoj medicini stalno rastu i sve su dostupniji. AIBDS će imati pristup i mogućnost konzultacije stotina ili tisuća medicinskih knjiga i znanstvenih radova, ali i drugih dijagnostički korisnih izvora, poput zbirki radioloških snimki, snimki magnetske rezonancije, uzoraka krvi i sl. Nijedan ljudski liječnik neće raspolagati tolikim znanjem i podacima, a osobito neće biti u stanju raspolagati njima gotovo trenutno, odnosno postavljati dijagnoze jednakom brzinom i pouzdanošću. Iako se AIBDS temelji na rezultatima različitih područja umjetne inteligencije, poput računalnog vida

[2] Za pregled medicinskih primjena umjetne inteligencije vidi npr. Eric J. TOPOL, High-Performance Medicine. The Convergence of Human and Artificial Intelligence, *Nature Medicine*, 25 (2019) 1, 44-56. Informativan osvrt na hrvatskom je Ivan SLADE-ŠILOVIĆ, Razvoj umjetne inteligencije u zdravstvu i zdravstvo sutrašnjice, *Medix*, 22 (2016) 121-122, 63-64.

[3] Za pregled etičkih aspekata medicinskih primjena umjetne inteligencije vidi npr. Alessandro BLASIME, Effy VAYENA, The Ethics of AI in Biomedical Research, Patient Care, and Public Health, u: Markus D. DUBBER, Frank PASQUALE, Sunit DAS (ur.), *The Oxford Handbook of Ethics of AI*, New York, Oxford University Press, 2020, 703-718.

[4] U upotrebi su i nazivi poput »Machine Learning Expert Systems« ili samo »Expert Diagnostic Systems«. Radi zadržavanja veze s međunarodno uvriježenom kraticom za umjetnu inteligenciju (AI) u radu se koristi kratica AIBDS.

(omogućujući, primjerice, uspoređivanje rendgenskih snimki) i procesiranja prirodnog jezika (radi, primjerice, izravne komunikacije s pacijentima), njihov ključan i najintrigantniji aspekt je strojno učenje (*machine learning*).

U najosnovnijim crtama: sustavi strojnog učenja nisu za rješavanje posebnih zadataka programirani po unaprijed strogo definiranim pravilima, već im je umjesto toga omogućen pristup određenom skupu podataka u kojemu sami otkrivaju korisne obrasce njihova rješavanja i tako sami *uče* na koje se sve načine može doći do rješenja. Budući da ovise o dostupnosti i sposobnosti obrade velikih skupova podataka, nagao razvoj različitih sustava strojnog učenja počeo je uvelike zahvaljujući razvoju interneta i znatnom povećanju komputacijske snage računala. Strojno učenje u mnogočemu već nadmašuje ljude i ima brojne komercijalne primjene, kao što su označavanje fotografija na društvenim mrežama, računalni sustavi za prevođenje, osobni asistenti na pametnim telefonima, sustavi preporuka u *online* trgovini, otkrivanje prijevara u poslovanju kreditnim karticama i sl.[5]

U medicini sustavi strojnog učenja nisu više novost. Više se ili manje već koriste kao pomagala liječnicima radi što preciznije analize raznih snimki prilikom dijagnosticiranja bolesti poput raka kože ili upale pluća. No, izgledna je i njihova šira upotreba. Primjerice, elektronički zdravstveni kartoni pacijenata povezat će se s raznim prenosivim uređajima (*wearables*) i njihovim senzorima, što će omogućiti njihovo dijagnostičko ažuriranje u stvarnom vremenu, a time i znatno ranije poduzimanje i individualnim osobama prilagođene intervencije (tzv. personalizirana medicina).[6]

Funkcioniranje AIBDS-a često se ilustrira analogijom s poznatim *online* platformama koje prodaju glazbu ili filmove. Takve platforme koriste strojno učenje da bi u podacima o velikom broju korisnika otkrile pravilnosti (poput sličnosti u ukusima) na temelju kojih za individualne korisnike predviđaju koji bi im se još glazbeni ili filmski naslovi mogli svidjeti (i prodati). Na sličan način AIBDS koristi strojno učenje da bi analizirao velike skupove medicinskih podataka i otkrivao ljudskom oku i umu teško uočljive pravilnosti (poput specifičnih simptoma u velikom broju sličnih ranijih slučajeva) te tako dolazio do vrlo pouzdanih dijagnoza i preporuka za liječenje.

AIBDS je u razvoju i potreban je oprez prilikom predviđanja kada će on – posebice ako bi trebalo postavljati dijagnoze bez ikakvog sudjelovanja ljudi liječnika – postati uobičajen dio medicinske svakodnevice. No, sudeći prema načinu i brzini kojom umjetna inteligencija ulazi u upotrebu u drugim područ-

[5] Izloženi prikaz strojnog učenja oslanja se na THE ROYAL SOCIETY, *Machine Learning. The Power and Promise of Computers that Learn by Example*, (04.2017), www.royalsociety.org/machine-learning (28.09.2020). U ovome radu, kao što je to čest slučaj u sličnim raspravama, termin »strojno učenje« koristim u relativno širokom smislu, tako da obuhvaća i metode i tehnike strojnog učenja poznate kao »duboko učenje« (*deep learning*).

[6] Usp. David S. WATSON i dr., Clinical Applications of Machine Learning Algorithms. Beyond the Black Box, *British Medical Journal*, 364 (2019) 1-4, 1; https://doi.org/10.1136/bmj.l886.

jima života, razložno je pretpostaviti da će se to isto u dogledno vrijeme dogoditi i s AIBDS-om i medicinom. Važno je naglasiti, dakako, da ključan razlog za njihovu širu primjenu, kao što je to i inače slučaj s takvim tehnologijama, neće biti to što će oni postati savršeni dijagnostičari koji nikada ne griješe, već to što će postati dijagnostičari koji, statistički gledano, griješe manje u usporedbi s ljudima dijagnostičarima. Neupitno je, međutim, da će njihov razvoj otvoriti i stanovita etička pitanja. Jedno od glavnih etičkih pitanja o AIBDS-u odnosi se na strojno učenje i tehnički problem poznatiji kao problem »transparentnosti«, problem »neobjašnjivosti« ili problem »crne kutije« (*black box*). On se obično definira ovako:

> »Nakon što su uvježbani (*trained*), mnogi sustavi strojnog učenja postaju 'crne kutije' čije su metode točne, ali ih je teško interpretirati. Iako takvi sustavi mogu proizvoditi statistički pouzdane rezultate, krajnji korisnik neće nužno biti u stanju objasniti kako su ti rezultati generirani ili koja su posebna svojstva danog slučaja bila važna za konačnu odluku.«[7]

Drugim riječima, problem je u tome što će za dani sustav strojnog učenja njegovi korisnici, ali možda i njegovi dizajneri, moći u najboljem slučaju znati koje su informacije u sustav ušle (*input*) i poslužile mu kao podaci za vježbu (*training data*) te koje su informacije iz sustava izašle kao rješenja ili odluke (*output*). Međutim, uslijed kompleksnosti sustava i količine podataka koje koristi, neće se moći znati na koji je način došao do tih rješenja ili odluka (sustavi strojnog učenja, prisjetimo se, nisu unaprijed programirani pravilima za rješavanje problema, već sami stvaraju takva pravila tijekom procesa »učenja« i dotadašnjeg rada). Povezani je problem što za određeni AIBDS, primjerice, neće uvijek biti moguće sa sigurnošću utvrditi je li u analiziranim podacima otkrio stvarne i dijagnostički relevantne uzročne veze ili tek neke zanimljive, ali uzročno i dijagnostički irelevantne pravilnosti (korelacije). Takav bi sustav, štoviše, mogao i propustiti otkriti neke očite i za zdravlje pacijenata iznimno važne uzročne veze.[8]

---

[7] The Royal Society, *Machine Learning…*, 93.

[8] Primjer kojim se često ilustrira ovaj drugi problem jest projekt s kraja 1990-ih kojim se htjelo procijeniti primjenjivost strojnog učenja za predviđanje smrtnosti za različite pacijente s upalom pluća. Jedan je određeni model strojnog učenja griješio s obzirom na pacijente koji su imali *i* upalu pluća *i* astmu. I dok bi ljudski dijagnostičar takve pacijente svrstao u visokorizičnu skupinu (koju treba odmah hospitalizirati), model ih je svrstao u niskorizičnu skupinu (koja se može liječiti i kod kuće). Model, samo na temelju podataka, nije shvatio da astmatičari s upalom pluća u pravilu primaju pojačanu medicinsku skrb i da je njihova stopa smrtnosti od upale pluća uslijed toga niža, a ne uslijed toga što je astma za upalu pluća irelevantna (usp. Rich CARUANA i dr., Intelligible Models for HealthCare. Predicting Pneumonia Risk and Hospital 30-day Readmission, *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, 1721-1730; https://doi.org/10.1145/2783258.2788613).

2. AIBDS kao prijetnja autonomiji pacijenata

Autonomija je jedno od osnovnih načela biomedicinske etike koje podrazumijeva, između ostaloga, pravo pacijenata da sami odlučuju hoće li i na koji način biti liječeni, ali i obvezu liječnika i ostalog medicinskog osoblja da to pravo i odluke poštuju. Inzistiranje na autonomiji pacijenata počinje u drugoj polovici 20. stoljeća kao odgovor na tradicionalni liječnički »paternalizam«: model odnosa liječnik-pacijent u kojemu je liječniku (kao stručnjaku) dopušteno pacijentu (kao nestručnjaku), radi njegova dobra, uskratiti važne informacije, pa čak i obmanjivati ga i donositi odluke umjesto njega. U konkretnoj praksi primjena načela autonomije u pravilu se odvija putem »informiranog pristanka«, procedure u kojoj pacijent daje izričit pristanak na predloženo liječenje na temelju njegova jasnog razumijevanja i shvaćanja njegovih mogućih, osobito negativnih, posljedica.[9]

Da bismo autonomiju i AIBDS smjestili u zajednički kontekst, zamislimo sljedeću situaciju: Otišli ste u bolnicu na pretrage koje obavlja AIBDS. Na temelju vaših medicinskih podataka te analize opsežne medicinske literature i tisuća sličnih slučajeva, AIBDS vam dijagnosticira bolest *x* i preporučuje što hitniji početak liječenja *y*. Budući da *y* može imati ozbiljnu nuspojavu *z*, prije pristajanja na liječenje htjeli biste znati koliko je izvjesno da imate *x* i koliko je izvjesna *z*. AIBDS, uslijed kompleksnosti svojih algoritama i količine korištenih podataka, ne može objasniti način kako je postavio dijagnozu. Liječnik koji opslužuje AIBDS može vam objasniti tek da je *x* doista opasno po život i da *y* jest jedno od standardnih liječenja, ali i da su dijagnoze AIBDS-a točne u čak 87 % slučajeva, dok su dijagnoze ljudskih liječnika točne u 67 % slučajeva. Iako nevoljko, pristajete na liječenje. Jeste li odluku donijeli autonomno, odnosno je li vaš pristanak bio informiran i moralno neproblematičan?

U literaturi, ali i u raznim dokumentima stručnih tijela i organizacija, rašireno je mišljenje da bi AIBDS, u situacijama poput opisane, bili prijetnja autonomiji pacijenata. Ističe se:

> »Budući da se sve više dijagnostičkih i terapijskih intervencija počinje temelji na strojnom učenju u medicini [*machine learning in medicine*, MLm], može biti potkopana autonomija pacijenata u procesima odlučivanja o njihovu zdravlju i mogućnost zajedničkog odlučivanja. To bi se dogodilo, primjerice, ako oslanjanje na automatizirana oruđa za odlučivanje smanji prigodu za smislen dijalog između pružatelja zdravstvene skrbi i pacijenata...«[10]

[9] Usp. Tom L. BEAUCHAMP, Ruth R. FADEN, Informed Consent, u: Stephen G. POST (ur.), *Encyclopedia of Bioethics*, New York, Thomson Gale – Macmillan, 2004, 1271.

[10] Effy VAYENA, Alessandro BLASIMME, I. Glenn COHEN, Machine Learning in Medicine. Addressing Ethical Challenges, *PLOS Medicine*, 15 (2018) 11, 1-4, 3; https://doi.org/10.1371/journal.pmed.1002689.

Slično se upozorava i na opasnosti algoritama strojnog učenja za proceduru informiranog pristanka:

> »Često spominjana prepreka široj kliničkoj upotrebi strojnog učenja nedostatak je razumijevanja kod pacijenata i liječnika o tome kako ono donosi svoja predviđanja. [...] Ako liječnici ne razumiju zašto je algoritam dao neku dijagnozu, zašto bi onda pacijenti vjerovali preporučenome tijeku liječenja? Je li informirani pristanak uopće moguć bez nekog poimanja o tome kako je model došao do svog zaključka?«[11]

Razmatrajući sve širu primjenu umjetne inteligencije u zdravstvenoj skrbi i istraživanjima, Vijeće za bioetiku Nuffield (*Nuffield Council on Bioethics*) ističe da takvi sustavi, ako se »koriste za postavljanje dijagnoze ili izradu plana liječenja, a zdravstveni djelatnik nije u stanju objasniti kako se do njih došlo«, mogu »ograničavati pravo pacijenata da slobodno i informirano donose odluke o svome zdravlju«.[12] Razrađujući etička načela za primjenu umjetne inteligencije u medicini, Kraljevski australski i novozelandski koledž radiologa u sličnom duhu upozorava da »strojno učenje i umjetna inteligencija mogu dovesti do rezultata koje je teško interpretirati ili ponoviti« i inzistira da prilikom njihove primjene »liječnik mora biti u stanju interpretirati osnovu pomoću koje se došlo do rezultata«, odnosno da mora moći »razumjeti i objasniti rezultat koji može utjecati na skrb o pacijentu«.[13] Da bismo donekle precizirali izložene tvrdnje, korisno je poći od u biomedicinskoj etici uvriježene analize autonomije prema kojoj ona ima (minimalno) tri komponente: (1) intencionalnost, (2) slobodu od ograničenjâ i (3) razumijevanje.

2.1. Intencionalnost

Intencionalnost, koja se naziva i »sposobnost za djelovanje« ili samo »djelatništvo« (*agency*), podrazumijeva svijest o sebi kao biću koje ima sposobnost »formirati namjere da bi se učinilo ono za što se vjeruje da će dovesti do željenog stanja stvari«.[14] Da bi se smatrao intencionalnim – za razliku od slučajnog ili nehotičnog – postupak »mora odgovarati djelatnikovoj koncepciji tog postupka, čak i ako planirani ishod možda neće biti onakav kako je zamišljen«.[15] Ljudi

[11] Watson i dr., *Clinical Applications...*, 2.
[12] NUFFIELD COUNCIL ON BIOETHICS, *Artificial Intelligence (AI) in Healthcare and Research*, (15.05.2018), https://www.nuffieldbioethics.org/publications/ai-in-healthcare-and-research (28.09.2020), 5.
[13] ROYAL AUSTRALIAN AND NEW ZEALAND COLLEGE OF RADIOLOGISTS, Ethical Principles for Artificial Intelligence in Medicine, (08.2019), 5, https://www.ranzcr.com/college/document-library/ethical-principles-for-ai-in-medicine (28.09.2020).
[14] Bruce L. MILLER, Autonomy, u: Stephen G. POST (ur.), *Encyclopedia of Bioethics*, New York, Thomson Gale – Macmillan, 2004, 246.
[15] Tom L. BEAUCHAMP, James F. CHILDRESS, *Principles of Biomedical Ethics*, New York – Oxford, Oxford University Press, 2013, 104.

su u načelu sposobni vagati svoje oprečne želje i formirati namjeru da djeluju na temelju onih za koje smatraju da će dovesti do nekog određenog ishoda. Za intencionalnost kao sastavni dio autonomije obično se smatra da ne može imati stupnjeve. Nešto smo učinili ili namjerno ili nenamjerno. Besmisleno je reći, primjerice, da smo nešto učinili 70 % namjerno, a 30 % nenamjerno. AIBDS, ukratko, nema u sebi ništa što bi dovodilo u pitanje ovako shvaćenu intencionalnost pacijenata. AIBDS uopće nije zamišljen da na bilo koji način utječe na mentalna stanja pacijenata, poput namjera i želja, i stoga njegove dijagnostičke metode ne kompromitiraju intencionalnost kao komponentu autonomije.

2.2. Sloboda od ograničenjâ

Prema uvjetu slobode od ograničenjâ, koji se ponekad naziva i »uvjet neovisnosti« ili »uvjet izostanka kontrolirajućih instanci«, ne smiju postojati »utjecaji koji tako kontroliraju ono što osoba čini da se ne može utvrditi da ona to želi učiniti«.[16] Uvjet propisuje da »osoba mora biti lišena kontrolâ koje provode bilo izvanjski izvori bilo unutarnja stanja koja osobu lišavaju samousmjerenosti (*self-directedness*)«.[17] Jednostavnije rečeno, riječ je o nepostojanju bilo kakve vrste prisile ili mentalnog stanja uslijed kojeg osoba ne bi slobodno kontrolirala svoje odluke i postupke (primjerice, kada odluke donosimo pod prijetnjom fizičkog nasilja, ucjene ili smrti, ili pak kada ih donosimo pod utjecajem droge, alkohola ili duševne bolesti). Za razliku od uvjeta intencionalnosti, smatra se da uvjet slobode od ograničenjâ može biti stupnjevit (primjerice, netko nas može pustiti da sami odlučimo hoćemo li nešto učiniti, može nas malo uporniје nagovarati da to učinimo ili nam može otvoreno prijetiti ne bismo li to učinili). Ograničavanje slobode ili kontrole nad vlastitim postupcima također ne izgleda kao nešto što je inherentno AIBDS-u o kojemu se raspravlja u ovom radu. U najgorem bi se slučaju moglo reći da nezavidan položaj – želja za spašavanjem vlastitog zdravlja ili života – pacijente čini sklonijim pristajanju na upotrebu AIBDS-a. No, valja uočiti, bez obzira na to što je riječ o sustavima »umjetne inteligencije«, da stvaranje te sklonosti samom AIBDS-u nije ništa više svojstveno nego drugim medicinskim tehnologijama.

2.3. Razumijevanje

Uvjet razumijevanja (ili »uvjet racionalnosti«) glasi da »postupak nije autonoman ako ga djelatnik primjereno ne razumije«, što se može dogoditi uslijed niza razloga, kao što su bolest, iracionalnost, nezrelost ili loša komunikacija.[18]

[16] Miller, *Autonomy…*, 246.
[17] Beauchamp, Childress, *Principles...*, 104-105.
[18] Usp. *isto*, 104.

Smatra se da razumijevanje – jednako kao i sloboda od ograničenjâ – jest stvar stupnja i da autonomija pacijenata ne zahtijeva puno razumijevanje relevantnih medicinskih postupaka, već samo njegovu »dostatnu (*substantial*) razinu«.[19] Stupnjevanje razumijevanja i zahtjev za »određenom razinom razumijevanja« dijelovi su zdravorazumskog pristupa poimanju autonomije ne samo u biomedicinskoj etici nego i općenito u ljudskom životu. Naime, mi često ne raspolažemo punim razumijevanjem mnogih važnih aspekata naših života (primjerice, svih važećih zakona ili načina na koji funkcionira sustav opskrbe strujom), ali, unatoč tome, ne smatramo da nam to narušava autonomiju jer o njima imamo dostatno razumijevanje. U kontekstu biomedicinske etike pitanje razumijevanja kao komponente autonomije očito je ključno za raspravu o AIBDS-u i može mu se pristupiti, kao što ćemo pokazati u nastavku, kako iz perspektive pacijenata tako iz perspektive liječnika.

3. Iz perspektive pacijenata

Iz perspektive pacijenata, kritika AIBDS-a kao prijetnje autonomiji mogla bi polaziti od razlike između gotovo neograničene kompleksnosti AIBDS-a s jedne strane i ograničene racionalnosti prosječnog pacijenta s druge. Razvojem suvremene medicine, pacijentu postaje sve teže razumjeti i od strane čovjeka liječnika postavljene dijagnoze i objašnjenja. Štoviše, medicina, poput mnogih drugih područja znanosti, postaje sve više specijalizirana i znanja stručnjaka za jedno njezino područje često su teško objašnjiva i stručnjacima za njezina druga područja, a kamoli pacijentima kao laicima. Budući da je za neke dijagnoze nerijetko potrebna i suradnja stručnjaka za više područja, čini se nerealističnim očekivati da će ih prosječan pacijent moći primjereno razumjeti. Ovoj rastućoj kompleksnosti ljudske medicine – kritika bi mogla glasiti – AIBDS će dodati nov sloj nerazumljivosti i netransparentnosti, uslijed čega će razumijevanje pacijenata postati praktično nemoguće. Budući da pacijenti neće biti u stanju razumjeti kako AIBDS dolazi do svojih dijagnoza, oni neće moći na temelju relevantnih informacija dati ili uskratiti pristanak za liječenje i time će im autonomija biti ozbiljno narušena.

Izloženu je kritiku, međutim, relativno jednostavno poljuljati – i to priznanjem razlike na kojoj ona počiva, ali uz njezino drukčije tumačenje. Mnogi pacijenti doista – bilo zbog kompleksnosti medicine bilo zbog vlastitog nedostatnog obrazovanja – ne mogu razumjeti ni mnoge dijagnoze i objašnjenja ljudi liječnika (dijagnoze u čije postavljanje nije uključena nikakva umjetna inteligencija). Ako se jaz nerazumijevanja uslijed specijalizacije medicine, kao što je spomenuto, širi između samih liječnika, on se nedvojbeno širi i između liječ-

[19] *Isto*.

nika i pacijenata, toliko da, iz perspektive prosječnog pacijenta, nema zamjetne razlike u nerazumljivosti »umjetno« (AIBDS) i »prirodno« (ljudski liječnici) postavljenih dijagnoza.[20] Pokušaj odbacivanja AIBDS-a samo iz perspektive pacijenata, drugim riječima, suočava se s problemom konzistentnosti: Ako nerazumljivost »ljudski« postavljenih dijagnoza ne izaziva ozbiljniju zabrinutost oko autonomije pacijenata, zašto bi je izazivala nerazumljivost »računalno« ili »strojno« postavljenih dijagnoza?[21]

Ne treba isključiti ni mogućnost da budući pacijenti, uslijed rastuće primjene umjetne inteligencije u medicini, neće ni cijeniti autonomiju jednako ili na isti način kao što su je pacijenti nekada cijenili. Možda će nove medicinske tehnologije izmijeniti ljudske stavove prema autonomiji na sličan kao što su digitalne tehnologije općenito (poput društvenih mreža) izmijenile ljudske stavove prema privatnosti. U svojem razmatranju hipotetičkog ekspertnog sustava za otkrivanje raka, Anderson i Anderson uvjereni su da bi samim pacijentima učinkovitost takvog sustava bila važnija od njegove transparentnosti:

> »Zamislimo da je AI program u širokoj upotrebi i da podiže stopu uspješnosti s 92 % na gotovo 100 %. Bi li nas toliko brinulo kako to radi? Među medicinskim osobljem zasigurno bi postojala znatiželja, ali bi li to bilo važno pacijentima – onima na čije živote utječe prisutnost ili odsutnost stanica raka? Mislimo da ne bi. To je stoga što posao koji ovaj program radi, za razliku od drugih AI programa, jest zapravo crn ili bijel. On pronalazi rak, ako on postoji. Sve do čega nam je stalo jest koliko je uspješan u uočavanju stanica raka. Ako bi postigao stopu uspješnosti od 100 %, a ljudi bi i dalje za tim daleko zaostajali, program bi mogao zamijeniti ljude prilikom obavljanja ovoga posla, oslobađajući ih da obavljaju druge poslove u kojima bi njihova stručnost bila presudna.«[22]

[20] U članku Computing Machinery and Intelligence (*Mind*, 59 [1950] 236, 433-460), Alan TURING predložio je svoj čuveni test – nazvavši ga »igra oponašanja« (*imitation game*) – za procjenu toga kada se može reći da je računalo postalo inteligentno poput čovjeka. Njegov je prijedlog, ukratko, sljedeći: Ako razgovaramo, razmjenjujući poruke putem zaslona, s čovjekom i s računalom, te ako, nakon određenog broja pitanja i odgovora, nismo u stanju razlučiti koji je od naših sugovornika računalo, a koji čovjek, onda je računalo prošlo test i ima istu vrstu inteligencije kao čovjek. Primijenimo li ideju »igre oponašanja« ili »Turingova testa« na AIBDS, teza bi mogla glasiti da prosječan pacijent vjerojatno ne bi ni uočio razliku između »ljudski« i »strojno« postavljenih dijagnoza.

[21] Slikovito rečeno, ako činjenica da pacijent ne razumije dijagnozu koju je na temelju svog znanja i iskustva zajednički postavilo troje ljudi liječnika, ne izaziva zabrinutost oko autonomije, zašto bi zabrinutost izazivala činjenica da pacijent ne razumije dijagnozu koju je postavio AIBDS objedinjujući znanja i iskustva – pohranjenim u raznim bazama podataka – troje, 33 ili 333 ljudi liječnika?

[22] Michael ANDERSON, Susan L. ANDERSON, How Should AI Be Developed, Validated, and Implemented in Patient Care, *AMA Journal of Ethics*, 21 (2019) 2, 125-130, 125-126; doi: 10.1001/amajethics.2019.125. Prema nekim istraživanjima, postoji visoka spremnost javnosti za prihvaćanje umjetne inteligencije i robotike u zdravstvenoj skrbi, a ključni čimbenici koji utječu na tu spremnost su brzina i točnost u dijagnosticiranju i liječenju [usp. *What Doctor? Why AI and Robotics Will Define New Health* (srpanj, 2017); https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health.html (29.11.2020)]. Dakako, pitanje

Iz perspektive pacijenata čak bi se moglo argumentirati da AIBDS ne samo da ne ugrožava njihovu autonomiju, nego je i povećava. Prisjetimo se analize prema kojoj autonomija uključuje minimalno tri komponente od kojih su dvije stupnjevane: sloboda od ograničenjâ i razumijevanje. S jedne strane, moglo bi se tvrditi da AIBDS pozitivno utječe na razumijevanje pacijentova stanja jer povećavaju količinu i brzinu obrade informacija potrebnih za postavljanje dijagnoze. Iako nije riječ o izravnom povećanju razumijevanja samog pacijenta, nego o povećanju razumijevanja općenito, možda čak razumijevanja na razini same medicine, ono može ipak, posredno, uroditi time da pacijent ima širi izbor informacija i više opcija za odlučivanje. S druge strane, očekuje se da će AIBDS biti lišen nedostataka koje liječnici od krvi i mesa znaju imati, poput umora i dekoncentracije, ali i mogućih predrasuda nastalih uslijed njihova odgoja i obrazovanja u nekom društvu i vremenu. AIBDS će utoliko manje od liječnika[23] biti sklon »utjecati« na pacijente ili ih »usmjeravati« prema nekim odlukama koje se tiču njihova zdravlja, čime će, u izvjesnom smislu, čuvati i njihovu autonomiju. S obzirom na složeni odnos između liječnika i pacijenata, isto je pitanje – narušava li AIBDS autonomiju pacijenata? – nužno razmotriti i iz perspektive liječnika.

4. Iz perspektive liječnika

Iz perspektive liječnika, AIBDS bi mogao izazivati zabrinutost zbog pitanja odgovornosti za moguće štete uslijed njihovih pogrešaka. S jedne strane, poznato je da su slični sustavi umjetne inteligencije u pravilu proizvodi rada brojnih stručnjaka te da, kada imaju štetne posljedice, njihove zakonske i moralne procjene nailaze na problem »distribuirane odgovornosti« ili »problem mnogih ruku«: situaciju u kojoj se za štetne učinke odgovornim može smatrati kolektiv koji ih je uzrokovao, ali ne i bilo kojeg njegova člana posebno.[24] S druge strane, unatoč ovoj tendenciji distribuiranja odgovornosti za štetne posljedice sličnih sustava, odgovornost liječnika za štete koje bi AIBDS mogao prouzročiti vjerojatno ne bi bila na sličan način raspršena. Odnos između pacijenta i liječnika jedinstven je i ne može se svesti na odnos između kupaca ili korisnika teh-

koje će zahtijevati daljnja istraživanja, pogotovo kad AIBDS bude ulazio u širu upotrebu, jest koliko će pacijenti cijeniti vrijednosti poput autonomije, a koliko učinkovitost takvih sustava.

[23] Raširen je stav da će sustavi umjetne inteligencije uvijek uključivati neke »sklonosti« ili »predrasude« (*bias*), naime, zato što će biti »trenirani« na podacima koji će sami možda sadržavati takve sklonosti ili predrasude. Vidi npr. Linda NORDLING, Mind the Gap, *Nature*, 573, 2019, 103-105, kao i više tehnički rad David DANKS, Alex John LONDON, Algorithmic Bias in Autonomous Systems, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 4691-4697; https://doi.org/10.24963/ijcai.2017/654.

[24] O »distribuciji odgovornosti« ili »problemu mnogih ruku« u inženjerstvu vidi Ibo VAN DE POEL, Lambèr ROYAKKERS, *Ethics, Technology, and Engineering. An Introduction*, Oxford, Wiley-Blackwell, 2011, pogl. 9.

nologija te njihovih dizajnera ili operatera. Liječnici prema pacijentima imaju posebne dužnosti, poput dužnosti pružiti im relevantne informacije o njihovu zdravlju i mogućnostima liječenja. *Kodeks medicinske etike i deontologije Hrvatske liječničke komore* ističe, primjerice, da liječnik treba

> »na prikladan način obavijestiti pacijenta i/ili zastupnika o dijagnostičkim postupcima i pretragama, njihovim rizicima i opasnostima te rezultatima, kao i svim mogućnostima liječenja i njihovim izgledima za uspjeh te mu primjereno pružiti potrebne obavijesti da bi pacijent mogao donijeti ispravne odluke o dijagnostičkom postupku i predloženom liječenju.«[25]

Ako ne bude u stanju rekonstruirati postupke pomoću kojih je AIBDS postavio dijagnozu, liječnik ih neće moći prikladno prenijeti pacijentu. Nastaje opasnost da će netransparentnost i neobjašnjivost inherentne AIBDS-u biti prenesene na odnos između liječnika i pacijenta te da će povjerenje među njima biti ozbiljno narušeno.[26] To pak ukazuje na važnost perspektive liječnika za očuvanje autonomije pacijenata ili, točnije, na usku povezanost između razumijevanja liječnika, autonomije pacijenata i povjerenja koje među njima mora postojati. Beil i suradnici tu povezanost formuliraju na sljedeći način:

> »Autonomno odlučivanje pacijenta ili surogatnog odlučitelja zahtijeva dostatno razumijevanje relevantnih medicinskih informacija, kao i procesa donošenja odluka unutar medicinske zajednice, poput poštivanja smjernica. [...] Međutim, nije realistično pretpostaviti da će ovi uvjeti biti ispunjeni u svakom slučaju. Stoga, povjerenje između pacijenata, surogatnih odlučitelja i liječnika još uvijek jest glavni stup odlučivanja u intenzivnoj njezi. [...] Kao čuvar (*gatekeeper*) informacija, medicinski stručnjak – bez obzira podrazumijeva li to čovjeka ili neku buduću implementaciju umjetne inteligencije – djeluje kao čuvar pacijentove autonomije.«[27]

Autonomija u smislu punog razumijevanja relevantnih medicinskih informacija, dakle, ideal je koji u praksi – kod stvarnih pacijenata – najčešće nije ispunjen te stoga odstupanje od tog ideala mora biti kompenzirano povjerenjem u neku instancu koja takvim razumijevanjem raspolaže: bilo da je riječ o

[25] *Kodeks medicinske etike i deontologije Hrvatske liječničke komore*, https://www.hlk.hr/kodeks-medicinske-etike-i-deontologije.aspx (28.09.2020) čl. 2, st. 8.

[26] Povjerenje nije jedini aspekt tog odnosa koji bi AIBDS mogao narušiti. Drugi i podjednako zanimljiv aspekt više je moralno-psihološke naravi. Naime, za štete koje će AIBDS izazvati pacijentima, uslijed problema »distribuirane odgovornosti« ili »problema mnogih ruku«, možda neće biti moguće pronaći jedinstvenog krivca (krivnja bi mogla biti raspršena, primjerice, na programere, dizajnere senzora, upravu bolnice itd.). Izgledno je, međutim, da će u takvim situacijama liječnici ostati primarne mete negativnih emocija oštećenih pacijenata (ili članova obitelji) i da bi stoga mogli imati razloga za suzdržanost prema upotrebi AIBDS. Riječ je o jednoj mogućoj varijaciji ideje »retribucijskog jaza« (*retribution gap*) o kojemu raspravlja John DANAHER, Robots, Law and the Retribution Gap, *Ethics and Information Technology*, 18 (2016) 4, 299-309.

[27] Michael BEIL i dr., Ethical Considerations about Artificial Intelligence for Prognostication in Intensive Care, *Intensive Care Medicine Experimental*, 7 (2019) 1-13, 5-6; https://doi.org/10.1186/s40635-019-0286-6.

individualnim liječnicima, zdravstvenom sustavu u cjelini, a možda i općenito medicini kao znanosti. Ako pak nitko neće imati takvo razumijevanje i neće biti moguće da bilo tko objasni kako se dolazi do posebnih dijagnoza, izgledno je da će važan dio tog povjerenja nestati, a s njime i važan dio autonomije pacijenata. Kao što primjećuje Boddington, iako ni individualni liječnici »često ne razumiju u potpunosti kako neko liječenje djeluje i mogu samo nepotpuno predvidjeti štetne učinke«, problem s umjetnom inteligencijom je u tome što bi ona »mogla donositi nepredvidive odluke koje ne možemo objasniti niti *načelno*«.[28] Perspektiva liječnika – uključujući ideal načelne objašnjivosti u okviru medicine – ključna je za očuvanje autonomije pacijenata jer se upravo od liječnika, gotovo po definiciji, očekuje da predstavljaju interpretativne karike u lancu koji povezuje nestručne pacijente i hiperstručni AIBDS.

Naznačenu konceptualnu povezanost autonomije, razumijevanja i povjerenja ne treba shvatiti kao konačnu presudu protiv, a još manje kao kraj rasprave o AIBDS-u kao prijetnji autonomiji. Primjerice, oštrica optužbe za kršenje autonomije mogla bi se otupiti ukazivanjem na to da je standard »načelne objašnjivosti« vjerojatno previsok. U svojem razmatranju *black box* medicine, London upozorava da

> »empirijska otkrića u medicini često imaju bolje epistemičko uporište od teorija koje bi ih mogle objašnjavati i da su ateorijske, asocijacionističke i netransparentne (*opaque*) odluke u medicini češće nego što to kritičari uviđaju«.[29]

Uspješne medicinske intervencije često su se temeljile na iskustvenim opažanjima koja nisu bila istovremeno praćena odgovarajućim razumijevanjem relevantnih uzročnih veza zbog kojih su te intervencije uopće bile uspješne i učinkovite.[30] Neizvjesnost i nepotpunost na poseban su način inherentne samoj medicini i stoga »općeniti zahtjev da sustavi strojnog učenja u medicini moraju biti objašnjivi ili interpretabilni nije utemeljen i potencijalno je štetan«.[31] Slikovito rečeno, ako smatramo dopustivim i za autonomiju pacijenata neproblematičnim da liječnici ponekad djeluju pod geslom »Znamo da funkcionira u praksi, ali ne znamo funkcionira li u teoriji«, zašto ne bismo to isto smatrali i kada je riječ o AIBDS-u – ili barem kada je riječ o onom AIBDS-u koji će biti statistički uspješniji od ljudi dijagnostičara?[32]

[28] Paula BODDINGTON, *Towards a Code of Ethics for Artificial Intelligence*, Cham, Springer, 2017, 62 (kurziv dodan).

[29] Alex John LONDON, Artificial Intelligence and Black-Box Medical Decisions. Accuracy versus Explainability, *Hastings Center Report*, 49 (2019) 1, 15-21, 15; https://doi.org/10.1002/hast.973.

[30] Neki od primjera kojima London ilustrira ovu tvrdnju su aspirin, koji se gotovo stotinu godina propisivao kao analgetik, iako nisu bili objašnjeni mehanizmi njegova djelovanja, te litij, koji se koristi kao stabilizator raspoloženja iako nije posve poznato kako to postiže (usp. *isto*, 17).

[31] *Isto*, 15.

[32] Time se ne sugerira da je rasuđivanje liječnika i AIBDS-a isto. Kao što ističe Nicholson Price II, *black-box* medicina nalikuje tradicionalnom liječničkom oslanjanju na iskustvo i intuiciju (primjerice, kada liječnik kaže: »Iskušao sam ovo na pacijentima poput vas, djelovalo je i stoga ću

Zaključak

Pitanju o AIBDS-u i njegovu potencijalnom kršenju autonomije pacijenata, kao što smo vidjeli, moguće je pristupiti na dva načina. S jedne strane, čini se nespornim da AIBDS prijete autonomiji pacijenata ako njegove dijagnoze nisu transparentne, odnosno ako samim liječnicima ostaju načelno nerazumljive, a pacijentima neobjašnjive (informirani pristanak nije moguć). S druge strane, optužbu za netransparentnost i kršenje autonomije može se smatrati pretjeranom, a obranu AIBDS-a graditi na činjenici da je autonomija stvar stupnja i da je teško reći gdje ona počinje ili prestaje, činjenici da su u povijesti medicine uspješne intervencije nerijetko prethodile njihovu teorijskom razumijevanju i objašnjenju, ali i realnoj mogućnosti da pacijenti autonomiju možda neće ni cijeniti onoliko koliko će cijeniti statističku uspješnost AIBDS-a.

U dokumentu s etičkim smjernicama za pouzdanu ili povjerenja vrijednu (*trustworthy*) umjetnu inteligenciju, Stručna skupina Europske komisije upozorava da »sustavi umjetne inteligencije za preporuku glazbe ne otvaraju ista etička pitanja kao i sustavi umjetne inteligencije koji predlažu kritično liječenje«.[33] To upozorenje nedvojbeno je na mjestu i treba ga imati u vidu prilikom razvijanja AIBDS-a i sličnih medicinskih tehnologija. Međutim, činjenicu da takve tehnologije otvaraju etička pitanja i izazivaju osjetljivost ne treba unaprijed protumačiti kao da su one nedopustive, osobito ne u svjetlu razumnog očekivanja da će one unaprijediti brigu o ljudskom zdravlju i spašavanje ljudskih života – dakako, ni ovo očekivanje ne treba protumačiti kao opravdanje za njihovu nekritičku i neprovjerenu primjenu.

Vam preporučiti isto«). No također upozorava da se *black-box* medicina oslanja na znatno širi skup informacija i metode njihove provjere koje nisu tipične za liječničko iskustvo i liječenje utemeljeno na intuicijama [W. NICHOLSON PRICE II, Black-Box Medicine, *Harvard Journal of Law and Technology*, 28 (2015) 2, 419-467, 430].

[33] STRUČNA SKUPINA NA VISOKOJ RAZINI O UMJETNOJ INTELIGENCIJI, *Etičke smjernice za pouzdanu umjetnu inteligenciju*, (08.04.2019), Europska komisija, Bruxelles; https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (28.09.2020), 7.

Tomislav Bracanović*

Artificial Intelligence, Medicine and Autonomy

Summary

The paper considers the question whether artificial intelligence-based diagnostic systems (AIBDS) pose a threat to patient autonomy as one of the central principles of biomedical ethics. The first part of the paper explains what AIBDS are and how they work, with an emphasis on machine learning technology and its property of non-transparency. The second part of the paper presents several paradigmatic views on AIBDS as a threat to patient autonomy, with reference to the standard analysis according to which this autonomy includes (at least) three components: intentionality, freedom from constraints and understanding. In the third part of the paper, the arguments pro et contra of the thesis about AIBDS as a threat to autonomy are considered from the perspective of patients, while in the fourth part of the paper they are considered from the perspective of physicians. The paper shows how complex the discussion of these issues can be and what argumentative tools can be made available to opposing parties.

Key words: artificial intelligence, autonomy, biomedical ethics, diagnostic systems, machine learning, non-transparency.

(*na engl. prev.* Tomislav Bracanović)

* Tomislav Bracanović, PhD, Senior Research Fellow, Institute of Philosophy, Zagreb; Address: Ulica grada Vukovara 54, HR-10000 Zagreb, Croatia; E-mail: tbracanovic@ifzg.hr.