ORIGINALNI NAUČNI RAD
DOI: 10.5937/reci2518080O
UDC: 81'366.596
004.81

orcid.org/0009-0000-3413-4152

Jelena J. Ostojić[1]
Beogradski institut za humanistiku i socijalna istraživanja i
Četvrta gimnazija u Beogradu

PROTIVČINJENIČKI KONDICIONALI I PROBLEM EKSPLIKABILNOSTI VEŠTAČKE INTELIGENCIJE

Sažetak: Problem eksplikabilnosti se odnosi na to da ljudi nedovoljno razumeju zašto veštačka inteligencija (AI) donosi određene odluke jer su im sami principi i procesi koji stoje iza tih odluka netransparentni. U ovom radu će biti predstavljen predlog nekolicine autora koji smatraju da bi se protivčinjenički kondicionali mogli koristiti za povećanje eksplikabilnosti tako što bi ukazivali na to u kojim protivčinjeničkim situacijama bi bila doneta drugačija odluka. Razumevanje razloga za donošenje određene odluke i kako bi se ta odluka mogla promeniti u različitim uslovima je ključno za izgradnju poverenja u modele AI. Protivčinjenička objašnjenja nude to i zato se smatra da ona mogu biti dragocena ukoliko bi postala opšteprihvaćena.

U postojećoj literaturi o protivčinjeničkim objašnjenjima, koja ionako nije opsežna budući da se radi o novoj temi, nisam naišla na radove koji uključuju pregled teorija kondicionala. Nekolicina autora ih pominje samo u par rečenica. Svrha ovog rada je da se na jednom mestu predstave neki etički problemi upotrebe AI, protivčinjenička objašnjenja i standardne i pragmatičke teorije protivčinjeničkih kondicionala.

Najpre ću navesti neke probleme upotrebe AI i kratak pregled o tome šta (ni)je preduzeto da se oni reše, zatim pregled standardnih i pragmatičkih teorija

[1] ost_jelena@yahoo.com

kondicionala i, najzad, rešenje problema eksplikabilnosti pomoću protivčinjeničkih kondicionala.

Ključne reči: protivčinjenički kondicionali, problem eksplikabilnosti veštačke inteligencije, etična veštačka inteligencija

NEKI ETIČKI PROBLEMI PRIMENE VEŠTAČKE INTELIGENCIJE

Definicija veštačke inteligencije

Ne postoji univerzalno prihvaćena definicija veštačke inteligencije, a definicija koja se najčešće koristi opisuje AI kao tehnologiju za kreiranje sistema koji mogu da simuliraju ljudsku inteligenciju. U tradicionalnom programiranju, softverski inženjer piše niz uputstava koja računar sledi, a sistemi veštačke inteligencije su kreirani da rade sa određenim nivoom autonomije. Sistem koji prosto koristi precizno definisana pravila ne može se smatrati sistemom veštačke inteligencije.

Mašinsko učenje je podoblast veštačke inteligencije koja se bavi ekstrakovanjem informacija iz podataka. Naziv "učenje" potiče od toga što mašinsko učenje simulira način na koji ljudi uče. Ovi sistemi su u stanju da bez programiranja "korak po korak", kako se radi u tradicionalnom programiranju, identifikuju pravilnosti u datim podacima i daju predviđanja. Za razliku od tradicionalnog programiranja gde softverski inženjer formuliše precizna uputstva za obradu ulaznih podataka, algoritmi mašinskog učenja se obučavaju na skupovima podataka i mogu samostalno da obrađuju podatke, prepoznaju pravilnosti i, na osnovu toga, vrše predviđanja.

Duboko učenje je podoblast mašinskog učenja i ide još korak dalje. Modeli dubokog učenja koriste algoritme koji su napravljeni po ugledu na strukturu i funkcionisanje ljudskog mozga. Oni su pogodni za složene probleme i ogromne skupove podataka. Duboko učenje ima širok spektar primena, kao što su prepoznavanje govora, procesuiranje prirodnog jezika (razumevanje, generisanje i prevođenje jezika), primena u zdravstvu (medicinska slika za otkrivanje bolesti), u finansijama (otkrivanje prevara, predviđanje cene akcija, i analiziranje finansijskih podataka), robotici, proizvodnji samovozećih vozila i dronova itd.

Etika i AI

Upotreba veštačke inteligencije otvara značajna etička pitanja. Neka od njih se odnose na privatnost korisnika. Iako analiza podataka može biti korisna, neophodno je da se obezbedi da se ti podaci ne koriste na načine koji nisu u najboljem interesu korisnika. Lični podaci se prikupljaju najčešće na dva načina – preuzimanjem podataka sa interneta, najčešće bez saglasnosti (tzv. *scraping*) ili prenamenom podataka o korisnicima (Solove, 2025: 26). Za upotrebu i čuvanje podataka je neophodna dozvola korisnika i moraju se sprovesti

najstrože sigurnosne mere da bi se sprečilo otkrivanje osetljivih podataka (Zhang at al. 2023: 11). AI omogućava prikupljanje dosad nezabeležene količine podataka koji se koriste za analiziranje i utvrđivanje međusobnih veza između njih. Da bi se osiguralo da sistemi veštačke inteligencije rade bezbedno, efikasno i pravedno, neophodno je da podaci budu dostupni iz različitih grupa stanovništva, da budu kvalitetni, dobro organizovani, odgovarajuće označeni i pouzdano dobijeni (Tschider, 2021: 114). Takođe se postavlja pitanje o ispravnosti primene sistema nadzora koji prikupljaju informacije o aktivnostima ljudi. Iako nadziranje može doprineti bezbednosti, ipak može predstavljati pretnju za privatnost nadziranih osoba (Remian, 2019: 24).

Zatim se postavlja pitanje o tome ko ima pravo vlasništva i pristupa podacima, kako oni treba da se analiziraju, tumače, čuvaju i dele (Vebritha, 2024: 316-320). *Big data* kompanije su ušle u mnoge sektore i postavlja se pitanje kako bi one mogle da koriste podatke koje su prikupile kroz interakciju s korisnicima. Postoji rizik da prikupljeni ili izvedeni podaci mogu biti prodati trećim licima, kao što su onlajn trgovci, osiguravajuće agencije ili poslodavci, koji mogu da koriste te podatke za sopstvene svrhe (Holmes & Porayska-Pomsta, 2023: 124).

Predrasude i diskriminacija su ključni problemi u upotrebi veštačke inteligencije. Sistemi AI su trenirani pomoću podataka koji mogu sadržati predrasude i diskriminaciju i oni ih mogu zadržati, ili čak produbiti (Tschider, 2021: 112). Algoritam koji koristi ovakve podatke može proizvesti podjednako nepravedne odluke i predviđanja (Varona, D., & Suárez, 2022: 7). Na primer, alati za prepoznavanje govora koji se koriste za procenu znanja jezika, često se obučavaju na glasovnim snimcima bez akcenta, što može dovesti do negativnih procena jezičkih veština migranata koji imaju akcenat ili pojedinaca sa poteškoćama u govoru. Slično tome, alati za prepoznavanje lica mogu se prvenstveno obučavati na slikama belaca, tako da bi mogli biti manje precizni kada se primenjuju na osobe drugih rasa. Otkrivanje predrasuda može biti problematično zbog nedostatka transparentnosti u algoritamskom procesu donošenja odluka i ljudskih odluka u vezi sa njegovom primenom (Holmes & Porayska-Pomsta, 2023: 128).

Sledeće pitanje koje treba da se reši je pitanje ko je odgovoran kada je reč o upotrebi AI (Kousa & Niemi, 2023: 274). Nije jasno ko bi trebalo da se smatra odgovornim kada nešto krene naopako – kompanije i organizacije koje koriste AI, kompanije koje kreiraju AI, softverski inženjeri ili neko drugi. Prema nekim autorima, postoji tzv. "jaz u odgovornosti" jer je ponašanje autonomnog sistema koji koristi složene algoritme previše samostalno i nepredvidivo da bi bilo koji od ljudskih aktera koji su učestvovali u njegovom razvoju mogli biti odgovorni za njega (Königs, 2022: 2 i Matthias, 2004: 36). Još neki rizici upotrebe AI su da može da izazove nezaposlenost i proširenje jaza između bogatih i siromašnih. Postoje, naravno, i manje kontroverzni problemi, koji su etički jasni, kao zloupotreba AI u kriminalu i zločinima protiv čovečnosti.

Problem transparentnosti se odnosi na poteškoće u razumevanju i tumačenju načina na koji sistemi veštačke inteligencije donose odluke. Sistemi AI su neka vrsta crnih kutija - čak ni softverski inženjeri AI ne razumeju u potpunosti šta se dešava unutar mašina AI. Ovaj nedostatak transparentnosti može biti problematičan jer otežava ljudima da prihvataju i razumeju odluke i preporuke koje donose sistemi veštačke inteligencije, što može ograničiti njihovu efikasnost. Pored toga, nedostatak transparentnosti takođe može otežati prepoznavanje predrasuda ili grešaka u procesu donošenja odluka sistema AI.

Problem eksplikabilnosti je usko vezan za problem transparentnosti i odnosi se na potrebu da se AI modeli učine razumljivima tako da i kreator i korisnik mogu da shvate funkcionisanje veštačke inteligencije. (Chamola at al, 2023: 79002) Opšta uredba o zaštiti podataka EU (EU General Data Protection Regulation (GDPR)) ne sadrži pravno obavezujuće pravo na eksplikabilnost (Wachter at al, 2017: 1). Ono je samo implicitno sadržano u članu 22 ove uredbe, prema kojem lice na koje se podaci odnose ima pravo da se na njega ne primenjuje odluka zasnovana isključivo na automatskoj obradi, uključujući i profilisanje, koja proizvodi pravne efekte koji se na njega odnose ili na sličan način značajno utiču na njega. Osim što pravo na objašnjenje koje bi otvorilo "crnu kutiju" i omogućilo uvid to kako algoritmi donose odluke nije pravno obavezujuće, postoje i drugi problem vezani za ostvarenje ovog prava. Recimo, objašnjenje funkcionisanja složenih algoritama modela AI za donošenje odluka i razloga za odlučivanje u specifičnim slučajevima je tehnički prejak zahtev jer često ni sami softverski inženjeri koji kreiraju ove modele nisu u stanju da objasne šta se tačno dešava unutar "crne kutije" (Kroll, 2016: 65). Osim toga, takva objašnjenja bi mogla biti nerazumljiva čime bi se dovela u pitanje njihova vrednost i korisnost. Takođe, kompanije koje kontrolišu podatke mogu da ne budu voljne da otkriju detalje o funkcionisanju algoritama da bi izbegle otkrivanje poslovnih tajni, da bi izbegle kršenje prava i sloboda drugih i njihove privatnosti i moguće manipulisanje modelima koji donose odluke (Burrell, 2016: 638 i Ford at al, 2016:21). O rešenju problema eksplikabilnosti će biti reči u poglavlju "Rešenje problema eksplikabilnosti pomoću protivčinjeničkih kondicionala".

Šta (ni)je urađeno da bi se izbegli rizici upotrebe veštačke inteligencije?

Formulisani su mnogobrojni etički vodiči za regulisanje AI, kao što su Etičke smernice Grupe eksperata Evropske komisije za veštačku inteligenciju (Ai, 2019: 1-9), Preporuka Saveta za veštačku inteligenciju (OECD Legal Instruments, 2019: 1-2). i Pekinški konsenzus o veštačkoj inteligenciji i obrazovanju (Ienca & Vayena, 2020: 1-25). Ove dokumente su izradile vladine agencije, privatne kompanije, akademske i istraživačke institucije, nevladine organizacije, neprofitne organizacije i dobrotvorne organizacije (Ienca & Vayena, 2020: 3-17).

Problem sa etičkim smernicama je što su one, praktično, dobrovoljne. Pored toga, etičke smernice su često napisane opštim jezikom, što može ograničiti njihov praktični uticaj.

Međutim, oblast veštačke inteligencije nije bez zakonskih ograničenja. Postojeći zakoni, kao što su *Univerzalna deklaracija o ljudskim pravima, Međunarodna konvencija o građanskim i političkim pravima, Međunarodna konvencija o eliminaciji svih oblika rasne diskriminacije, Konvencija o eliminaciji svih oblika diskriminacije žena, Konvencija za zaštitu ljudskih prava i osnovnih sloboda, Povelja EU o osnovnim pravima unije,* itd. pružaju širok okvir za regulisanje AI. Međutim, ovi zakoni nisu napisani imajući na umu AI i stoga ne daju pravila specifična za njenu upotrebu. Dok trenutni okvir zasnovan na ljudskim pravima, demokratiji i vladavini zakona pruža odgovarajući kontekst, potreban je poseban obavezujući dokument za regulisanje AI. Veštačka inteligencija otvara probleme koji prevazilaze postojeće zakone, kao što su nezaposlenost, odnosi između ljudi i mašina, pristrasni algoritmi i opasnosti koje nosi buduća super-inteligencija. Da bi se pozabavila ovim izazovima, Evropska komisija je 2021. godine predložila *Zakon o veštačkoj inteligenciji* (*AI Act*). (Ostojic, 2024: 75-83). Kasnije će biti reči o jednom od mogućih načina rešavanja problema eksplikabilnosti pomoću protivčinjeničkih kondicionala.

PROTIVČINJENIČKI KONDICIONALI I EKSPLIKABILNOST

Definicija kondicionala

Kondicionali su izraženi rečenicama koje u gramatici nazivamo kondicionalnim ili pogodbenim, tj. složenim rečenicama oblika „Ako *A*, onda *B*“, ili „Da *A*, onda bi *B*“, ili „Kada bi *A*, onda bi *B*“, ili „Pošto *A*, onda *B*“, ili „Da je bilo *A*, bilo bi *B*“.Iskaz kojim se izražava uslov i koji obično stoji iza *ako, da, kada bi, pošto,* nazivamo *antecedensom* i označen je sa *A*; a drugi iskaz nazivamo *konsekvensom* (*B*).

Standardne teorije protivčinjeničkih kondicionala

Nakon Kripkeovog (Saul Kripke) pronalaska semantike mogućih svetova za modalne logike (Kripke, 1963: 67-96), dobijen je moćan jezik koji je mogao da se upotrebi u analizi mnogih pojmova. Ubrzo je Robert Stalnaker (Robert Stalnaker) upotrebio taj jezik da napravi semantiku za kondicionale tako što je modalnoj logici dodao pojam *funkcije selekcije* čiji je zadatak rangiranje mogućih svetova po tome koliko liče na aktualni svet. Po Stalnakeru, $A \mathrel{\square\!\!\rightarrow} C$[2] je istinito u aktuelnom svetu ako i samo ako je *C* istinito u najbližem svetu u kom je istinit antecedens. Stalnakerova ideja da se primeni semantika mogućih svetova za analizu kondicionala je prihvaćena i danas je preovladavajući način bavljenja kondicionalima.

Stalnakerova teorija je prva od tzv. *standardnih* teorija kondicionala, a sledeći predstavnik standardnih teorija kondicionala koga ćemo pomenuti je Dejvid Luis (David

[2] Ovo je Luisov znak za protivčinjenički kondicional.

Lewis). Ova dva filozofa su najvažniji predstavnici standardnih teorija i u literaturi se ovakav pristup često naziva Stalnaker-Luisovom teorijom.

U pomenutom tekstu, Stalnaker se bavi logičkim problemom kondicionala, tj, opisuje formalna svojstva kondicionalne funkcije. Ta funkcija je predstavljena sa „ako, onda“ i ona kao argument uzima uređeni par iskaza, a kao vrednost dobija opet iskaz. Na primer, od iskaza:

Kinezi su se umešali u rat u Vijetnamu.

i iskaza:

SAD su upotrebile atomsko oružje u ratu u Vijetnamu.

kao vrednost kondicionalne funkcije dobićemo iskaz:

Da su se Kinezi umešali u rat u Vijetnamu, SAD bi upotrebile atomsko oružje. (Stalnaker, 1968: 100).

Da bi nam približio intuiciju o tome kako u praksi procenjujemo istinitost kondicionala, Stalnaker se poziva na *Remzijev test* (Frank Ramsey). Remzijev test možemo predstaviti ovako: kondicional procenjujemo tako što najpre hipotetički dodamo antecedens datog kondicionala u skup naših verovanja, zatim izvršimo sva neophodna usaglašavanja neophodna za očuvanje konzistentnosti i na kraju razmatramo da li tada prihvatamo konsekvens. Ako prihvatamo, kondicional se smatra istinitim. Da bi Stalnaker formulisao teoriju, potrebno je da izvrši prelaz od uslova verovanja o kojima govori Remzijev test na istinitosne uslove kondicionala. Pojam mogućeg sveta mu omogućava ovaj prelaz. Mogući svet je „ontološki analogan skupu hipotetičkih verovanja“ (Stalnaker, 1968: 102). Mogući svet jednostavno predstavlja alternativu aktuelnom svetu. Stvari su mogle biti drugačije nego što jesu i pojam mogućeg sveta služi da se izrazi ta mogućnost. Mogućih svetova ima beskonačno mnogo, a nama je za ovu svrhu potreban onaj koji je po svemu sličan našem svetu osim po tome što je u njemu je *A* tačno.

Pogledajmo na Luisovom primeru kojim započinje knjigu *Counterfactuals* (Lewis, 1973: 5) kako se vrednuju kondicionali. Kada procenjujemo istinitosnu vrednost protivčinjeničkog kondicionala: „Da kenguri nemaju repove, prevrnuli bi se “, prema Luisu (i Stalnakeru, kao što smo videli), ukoliko se kenguri prevću u najbližim svetovima (kod Stalnakera u najbližem *svetu*) u kojima nemaju repove, onda je kondicional istinit. Najbliži mogući svet se minimalno razlikuje od aktuelnog, tek toliko koliko je neophodno za istinitost antecedensa. U ovom primeru, svetovi koje je izdvojila funkcija selekcije nalikuju na aktuelni svet po svemu osim po tome što u njima kenguri nemaju repove (i, eventualno, još po onim svojstvima koja slede iz toga). Kondicional „Da je bilo *A*, bilo bi *B*“ je tačan (lažan) ako samo ako je *B* tačno (lažno) u tom mogućem svetu. (Ostojić, 2016: 3-16).

Pragmatičke teorije protivčinjeničkih kondicionala

Druga važna grupa teorija kondicionala naziva se *pragmatičkim* teorijama kondicionala. Neki od zastupnika pragmatičkog pristupa su Kenet Varmbrod (Warmbrōd, 1981), Krispin Rajt (Wright, 1983), Džonatan Lou (Lowe, 1990), Kai fon Fintel (von Fintel, 2001), Entoni Gilis (Gillies, 2007) (koji primenjuju dinamičku semantiku), Vladan Đorđević (Djordjevic, 2005), Brit Brogard, Džo Salerno (Brogaard & Salerno, 2008) i dr. Pragmatičke teorije takođe koriste semantiku mogućih svetova ali, prema ovim teorijama, kondicional je striktna implikacija gde se skup dostiživih svetova menja s kontekstom i stvar je pragmatike kako će se taj skup odrediti. Bitna razlika između standardnih i pragmatičkih teorija je vezana za način shvatanja uloge konteksta u vrednovanju kondicionala. Prema Luisu, kontekst određuje poredak svetova po sličnosti u odnosu na bazični svet i jednom utvrđen redosled svetova ne menja tokom konverzacije. Ako se to desi, on to naziva promenom konteksta i prethodno vrednovani kondicionali u određenom razgovoru se moraju ponovo vrednovati s obzirom na novu funkciju selekcije ili relaciju sličnosti. Najnovije pragmatičke teorije usvajaju tzv. dinamički pristup. One dopuštaju izvesne promene konteksta a da se ne zahteva ponovno vrednovanje prethodnih kondicionala. (von Fintel, 2001 i Gillies, 2007) U specifičnosti raličitih pragmatičkih teorija se ne možemo upuštati ovde, samo bih naglasila da je prednost pragmatičkih teorija to što, za razliku od standardnih, dopuštaju zaključivanje iz kondicionalnih premisa pomoću tranzitivnosti, kontrapozicije i jačanja antecedensa pod određenim ograničenjima vezanim za promenu konteksta, što je u skladu s našom intuicijom i onim kako kondicionale koristimo u govoru. (Ostojić, 2016: 33-45).

Rešenje problema eksplikabilnosti pomoću protivčinjeničkih kondicionala

U ovom poglavlju će biti prikazano jedno moguće rešenje ovog problema. Objašnjenje odluka dobijenih isključivo na automatskoj obradi podataka zapravo i ne mora da zavisi od opšteg razumevanja načina na koji funkcionišu algoritamski sistemi, kako je navedeno u poglavlju o etičkim problemima AI da neki autori smatraju. Modeli AI mogu pružiti objašnjenja i bez zavirivanja u "crne kutije" i otkrivanja njene unutrašnjosti, tako što bismo koristili protivčinjenička objašnjenja (Wachter at al, 2017: 4). Da bi mašinsko učenje bilo interpretabilno, potrebno je da nam budu razumljivi razlozi koji stoje iza predviđanja i odluka modela, tj, da postoje objašnjenja ishoda modela.

U poslednje vreme, više autora smatra da su protivčinjenički kondicionali dobri kandidati za ovu svrhu. Rad u kom je prvi put predložena upotreba protivčinjeničkih kondicionala u ovu svrhu je Wachter at al. 2017, ali ona još nema široku primenu (Verma at al, 2021: 1). Protivčinjeničko objašnjenje otkriva šta je trebalo da bude drugačije u nekom slučaju da bi se dobio drugačiji ishod (Guidotti, 2024: 2771). Protivčinjenički kondicional pruža povratnu informaciju oblika „da je ulazni podatak (input) bio x` umesto x, onda bi izlazni podatak (output) modela mašnskog učenja bio y` umesto y." (Verma at al, 2021: 1).

Tako kondicionali omogućavaju objašnjenja koja imaju sledeću formu: Odbijen je Vaš zahtev za kredit zato što je Vaš prosečni mesečni prihod bio 100.000 dinara. Da je Vaš prihod bio 150.000 dinara, kredit bi Vam bio odobren. Vidimo da je nakon odluke o odbijanju kredita naveden protivčinjenički kondicional koji kaže kakav je svet morao biti da bi do željenog ishoda došlo. Rečeno terminologijom standardnih teorija kondicionala, u najbližem mogućem svetu u kojem podnosilac zahteva za kredit ima primanja od 150.000, on dobija kredit. Taj mogući svet se od aktuelnog razlikuje samo utoliko što u njemu ta osoba ima viša primanja. Naravno, moguće je da objašnjenje sadrži više od jednog kondicionala jer može postojati više poželjnih ishoda i moguće je da postoji više načina na koje se oni mogu postići.

U postojećoj literaturi, pojam objašnjenja se obično odnosi na pokušaj da se pronikne u unutrašnje stanje modela ili u logiku algoritma koja se koristi u procesu automatizovanog odlučivanja. Nasuprot tome, protivčinjenički kondicionali navode na koji način je odlučivanje zavisilo od spoljnih faktora. Ova razlika je važna zato što se u mašinskom učenju unutrašnje stanje algoritama može sastojati od miliona međusobno povezanih promenljivih. Protivčinjenička objašnjenja su tako formulisana da proužaju minimalnu količinu informacija koje su potrebne da bi odluka bila drugačija i ne zahtevaju razumevanje unutrašnje logike modela (Wachter at al, 2017: 13).

Još jedna prednost ovakvog tipa objašnjenja je u tome što ona pružaju informaciju o razlozima iz kojih je neka odluka donesena. Ta objašnjenja mogu obezbediti argumente na osnovu kojih bi se izvesna odluka mogla osporiti i omogućavaju uvid u to šta bi se moglo promeniti da bi se u budućnosti postigao željeni ishod.

Zaključak

U radu su najpre navedeni neki etički problemi upotrebe AI. Jedan od problema na putu ka izgradnji poverenja u modele AI je problem eksplikabilnosti AI. On se odnosi se na to da ljudi nedovoljno razumeju zašto AI donosi određene odluke jer su im sami principi i procesi koji stoje iza tih odluka uglavnom nejasni. U radu je izloženo jedno od predloženih rešenja ovog problema. U radu je takođe dat i pregled standardnih i pragmatičkih teorija kondicionala zato što nekoliko autora smatra da su protivčinjenički kondicinali alat koji nam može obezbediti objašnjenja za konkretne odluke modela AI. Oni bi se mogli koristiti za povećanje eksplikabilnosti tako što bi ukazivali na to u kojim protivčinjeničkim situacijama bi bila doneta drugačija odluka. Razumevanje razloga za donošenje određene odluke i kako bi se ta odluka mogla promeniti u različitim uslovima je ključno za izgradnju poverenja u modele AI. Na ovaj način modeli AI mogu pružiti objašnjenja i bez zavirivanja u “crnu kutiju”, tj, unutrašnje stanje modela ili u logiku algoritma koja se formirala u procesu automatizovanog odlučivanja.

Literatura

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*(3), 431-440. Dostupno preko: https://link.springer.com/article/10.1007/s43681-021-00096-7 [15.3.2024]

Ai, H. (2019). High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, *6*. 15.3.2024. Dostupno preko: https://www.aepd.es/sites/default/files/2019-09/ai-definition.pdf [15.3.2024]

Brogaard, B., & Salerno, J. (2008). Counterfactuals and context. *Analysis*, 68(1), 39-46. Dostupno preko: https://www.jstor.org/stable/25597849 [8.7.2025]

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1). Dostupno preko: https://journals.sagepub.com/doi/pdf/10.1177/2053951715622512 [15.3.2024]

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (xai). *IEEe Access*, 11, 78994-79015. Dostupno preko: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10188681 [8.7.2025]

Ienca, M., & Vayena, E. (2020). AI Ethics Guidelines: European and Global Perspectives-Provisional report. *Report No. CAHAI* (2020). Dostupno preko: https://rm.coe.int/cahai-2020-07-fin-en-report-ienca-vayena/16809eccac [15.3.2024]

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nature neuroscience*, *8*(9), 1255-1262. Dostupno preko: https://pure.mpg.de/rest/items/item_2615172_3/component/file_2622686/content [15.3.2024]

Djordjevic V. (2005) *Counterfactuals* (Doctoral dissertation, University of Alberta)

Ford, R. A., Price, W., & Nicholson, I. I. (2016). Privacy and accountability in black-box medicine. *Mich. Telecomm. & Tech. L. Rev.*, *23*, 1. Dostupno preko: https://heinonline.org/HOL/LandingPage?handle=hein.journals/mttlr23&div=5&id=&page= [15.3.2024]

Gillies, A. S. (2007). Counterfactual scorekeeping. *Linguistics and Philosophy*, 30, 329-360. Dostupno preko: http://web.mit.edu/fintel/fintel-gillies-ose2.pdf [8.7.2025]

Guidotti, R. (2024). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5), 2770-2824. Dostupno preko: https://link.springer.com/content/pdf/10.1007/s10618-022-00831-6.pdf [8.7.2025]

Holmes, W., & Porayska-Pomsta, K. (2023). The ethics of artificial intelligence in education. *Lontoo: Routledge*.

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, 24(3), 36. Dostupno preko: https://link.springer.com/content/pdf/10.1007/s10676-022-09643-0.pdf [8.7.2025]

Kousa, P., & Niemi, H. (2023). Artificial intelligence ethics from the perspective of educational technology companies and schools. *Learning: Designing the Future*, *283*. Dostupno preko: https://library.oapen.org/bitstream/handle/20.500.12657/60151/1/978-3-031-09687-7.pdf#page=294 [15.3.2024]

Kripke, S. A. (1963). Semantical analysis of modal logic i normal modal propositional calculi. *Mathematical Logic Quarterly*, 9(5-6), 67-96.

Kroll, J. A. (2015). *Accountable algorithms* (Doctoral dissertation, Princeton University).

Lewis, D. (1973) *Counterfactuals*, Oxford: Blackwell.

Li, Z. (2024). Ethical frontiers in artificial intelligence: navigating the complexities of bias, privacy, and accountability. *International Journal of Engineering and Management Research*, 14(3), 109-116. Dostupno preko: https://ijemr.vandanapublications.com/index.php/j/article/view/1610/1482 [8.7.2025]

Lowe, E. J. (1990). Conditionals, context, and transitivity. *Analysis*, 50(2), 80-87. Dostupno preko: https://www.jstor.org/stable/3328851 [8.7.2025]

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6, 175-183.

OECD Legal Instruments. (2019). Recommendation of the Council on Artificial Intelligence. *Organization for Economic Cooperation and Development*. Paris: OECD LI

Ostojic, J. (2024) Ethics of Artificial Intelligence. Marinković, P. Krstić (ur.) Education in *Humanism, Posthumanism, Anti-Humanism: Educational Perspectives* (75-83). London: TransNational Press London.

Ostojić, J. (2015). *Kondicionali, kontekst i znanje* (Doctoral dissertation, University of Belgrade (Serbia)).

Remian, D. (2019) Augmenting Education: Ethical Considerations for Incorporating Artificial Intelligence in Education. *Instructional Design Capstones Collection. 52.* Boston: University of Massachusetts Boston

Solove, D. J. (2025). Artificial intelligence and privacy. *Fla. L. Rev.*, 77, 1. Dostupno preko: https://scholarship.law.ufl.edu/cgi/viewcontent.cgi?article=4187&context=flr [8.7.2025]

Stalnaker, R. C. (1968). A theory of conditionals. Harper, Stalnaker i Pearce (ur.), *Ifs: Conditionals, belief, decision, chance and time* (41-55). Dordrecht: Springer Netherlands.

Tschider, C. A. (2021). AI's Legitimate Interest: Towards a public benefit privacy model. *Hous. J. Health L. & Pol'y*, 21, 125. Dostupno preko: https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3725933 [8.7.2025]

Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). Counterfactual thinking: an fMRI study on changing the past for a better future. *Social cognitive and affective neuroscience*, *8*(5), 556-564. Dostupno preko: https://academic.oup.com/scan/article/8/5/556/1676111 [15.6.2025]

Varona, D., & Suárez, J. L. (2022). Discrimination, bias, fairness, and trustworthy AI. *Applied Sciences*, 12(12), 5826. Dostupno preko: https://www.mdpi.com/2076-3417/12/12/5826 [8.7.2025]

Vebritha, S. (2024). Redefining Ownership and Originality in the Age of AI: A Legal and Ethical Review. *Sinergi International Journal of Law*, 2(4), 312-324. Dostupno preko: https://journal.sinergi.or.id/index.php/law/article/view/726/539 [8.7.2025]

Verma, S., Dickerson, J., & Hines, K. (2021). Counterfactual explanations for machine learning: Challenges revisited. *arXiv preprint arXiv:2106.07756*. Dostupno preko: https://arxiv.org/pdf/2106.07756 [8.7.2025]

Von Fintel Kai (2001) Counterfactuals in a dynamic context. in A *Life in Language*, priredio Ken Hale, Cambridge (MA)

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841. Dostupno preko: https://arxiv.org/pdf/1711.00399 [15.6.2025]

Warmbrōd, K. (1981). Counterfactuals and substitution of equivalent antecedents. *Journal of Philosophical Logic*, 267-289. Dostupno preko: https://www.jstor.org/stable/30227193 [8.7.2025]

Wright, C. (1983). Keeping track of Nozick. *Analysis*, 43(3), 134-140. Dostupno preko: https://www.jstor.org/stable/3327431 [8.7.2025]

Zhang, Y., Wu, M., Tian, G. Y., Zhang, G., & Lu, J. (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics. *Knowledge-Based Systems*, *222*, 106994. Dostupno preko: https://www.mdpi.com/1424-8220/23/3/1151 [8.7.2025]

Jelena J. Ostojić

COUNTERFACTUALS AND ARTIFICIAL INTELLIGENCE EXPLICABILITY PROBLEM

Summary: In the first part of the text, I present the most important ethical problems in the application of AI, such as perpetuating existing prejudices and discrimination and increasing injustice and inequality, explainability, privacy, transparency, responsibility, and autonomy. The problem of explainability refers to the fact that people do not sufficiently understand why AI makes certain decisions because the principles and processes behind those decisions are not sufficiently clear to them. Recently, several authors suggested that counterfactual conditionals could be used to increase explainability. The second part is a review of the most important theories of counterfactuals.

The third part is a review of the application of counterfactuals in explanations. Understanding the reasons for making particular decisions and why that decision might change under different conditions is the key to building trust in AI models. Explanations can serve many purposes: to inform and help the user understand why a particular decision was made, to provide grounds to contest adverse decisions, and to understand what can be changed to achieve a desired result in the future. In the current literature, "explanation" includes opening the "black box" to provide insight into the internal decision-making process of algorithms. However, explaining the functionality of complex algorithmic decision-making models and their rationale in specific cases is a technically challenging problem. In contrast, counterfactual explanations describe dependency on the external facts that led to that decision. Thus, they can, in principle, be offered without opening the "black box", what is seen as their main advantage.

Keywords: counterfactual conditionals, the artificial intelligence explicability problem, ethical artificial intelligence

Datum prijema: 31.8.2025.

Datum ispravki: 30.10.2025.

Datum odobrenja: 3.11.2025.