Predrasude ugrađene u algoritme i automatizirane procese

Talan, Đanino

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2022./ 2023.

Đanino Talan

Predrasude ugrađene u algoritme i automatizirane procese

Završni rad

Mentor: dr.sc. Denis Kos, doc.

Zagreb, kolovoz 2023.

Izjava o akademskoj čestitosti

Izjavljujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

SADRŽAJ

Sadržaj

1. UVOD

U suvremenom tehnološkom napretku, umjetna inteligencija (dalje: AI) predstavlja izvanredno postignuće, pokazujući izvanrednu sposobnost strojeva da oponašaju ljudsku inteligenciju. Kako se sustavi umjetne inteligencije sve više integriraju u različite aspekte društva, od zdravstva i financija do zabave i upravljanja, oni imaju potencijal da revolucioniraju način na koji radimo, komuniciramo i krećemo se svijetom. Međutim, ova brza integracija također iznosi na vidjelo izazov koji zahtijeva duboko istraživanje: raskrižje umjetne inteligencije, ugrađenih predrasuda i automatiziranih procesa. AI, koju karakterizira sposobnost učenja iz podataka i samostalnog donošenja odluka, nije imuna na nasljeđivanje pristranosti prisutnih u podacima iz kojih uči. Ovaj fenomen, nazvan "ugrađene predrasude", naglašava nenamjerno širenje društvenih pristranosti i predrasuda kroz sustave umjetne inteligencije. Posljedica je održavanje povijesnih nepravdi i nejednakosti u vremenu koje teži pravednosti i inkluzivnosti. Paralelno, val automatizacije pokretane umjetnom inteligencijom mijenja krajolik rada i produktivnosti. Automatizirani procesi imaju potencijal za povećanje učinkovitosti, smanjenje pogrešaka i kataliziranje inovacija. Ipak, kako ovi automatizirani sustavi preuzimaju uloge koje su nekoć imali ljudi, postavljaju se pitanja o etičkim implikacijama, potencijalnom premještanju poslova i potrebi za ljudskim nadzorom kako bi se osiguralo odgovorno donošenje odluka.

Ovaj rad pregled je zamršene korelacije između umjetne inteligencije, ugrađenih predrasuda i automatiziranih procesa. Udubljujući se u mehanizme koji pokreću te fenomene, analizirajući studije slučaja iz stvarnog svijeta i procjenjujući etičke i društvene implikacije, ovo istraživanje nastoji rasvijetliti izazove i prilike na ovom području. U konačnici, kritičkim ispitivanjem ovih isprepletenih aspekata, cilj nam je pridonijeti odgovornom razvoju i implementaciji AI tehnologija koje su usklađene s načelima jednakosti, pravednosti i dobrobiti ljudi.

Rad se sastoji od pet poglavlja. U uvodnom poglavlju prikazuje se predmet i cilj rada te njegova struktura. Drugo poglavlje obrađuje nastanak, napredak i primjenu umjetne inteligencije. Treće poglavlje obrađuje predrasude ugrađene unutar algoritama i automatiziranih procesa. Ovdje se istražuje problematika i pristranost automatizacije kao i algoritamska averzija. U četvrtom poglavlju istražuju se uzroci predrasuda i pristranosti, strategije detekcije predrasuda i ugrađenih pristranosti kao i primjeri predrasuda i pristranosti ugrađenih u algoritme i automatizirane procese. U završnom poglavlju izvode se zaključci na temelju istraživanja.

2. UMJETNA INTELIGENCIJA

Kako bi bilo moguće raspravljati o primjeni, kontroli i etici umjetne inteligencije, potrebno je prvo definirati pojam. Unatoč mnogim izvorima s različitim definicijama, postoje osnovna načela i ideje prema kojima se može doći do univerzalne definicije.

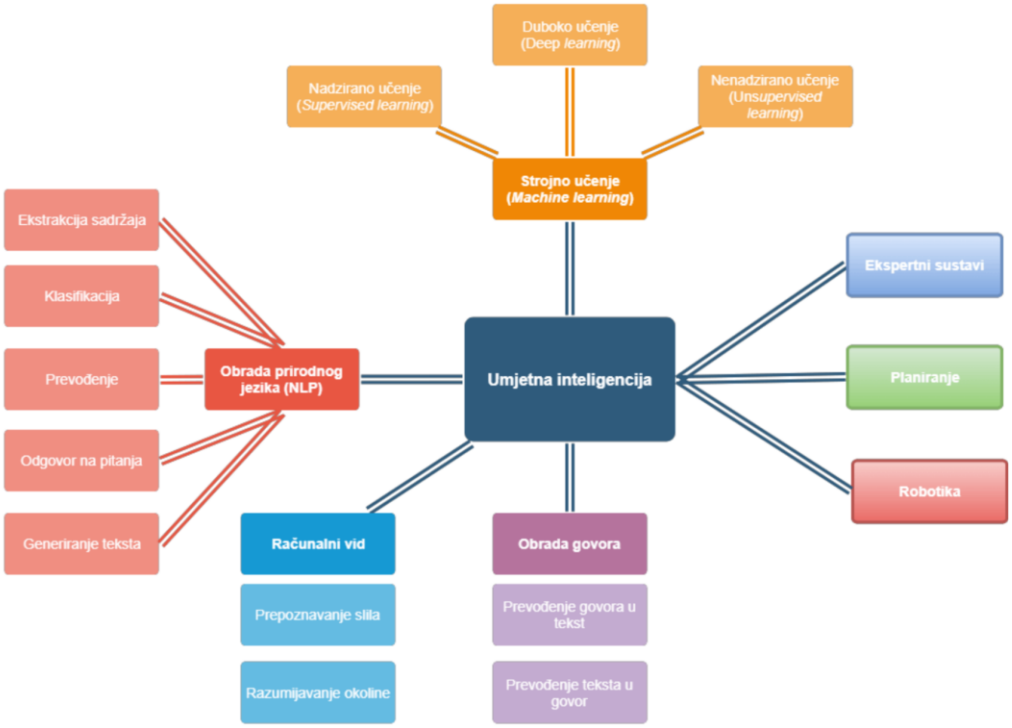
2.1. Definicija umjetne inteligencije

Umjetna inteligencija se može definirati kao određena skupina teorija kao i tehnika koje se upotrebljavaju za kreiranje strojeva koji su sposobni simulirati inteligenciju. Umjetna inteligencija kao takva predstavlja opći pojam koji kao takav podrazumijeva upotrebu računala za područje modeliranja inteligentnog ponašanja sa što manjom potrebom za ljudskim posredovanjem (Wamba-Tagumdje i sur., 2020). Umjetna inteligencija se ujedno odnosi na mogućnost stjecanja i primjenjivanja razno raznih vještina i znanja kako bi došlo do rješavanja određenog problema (Shabbir i Anwer, 2018). Za razliku od drugih tehnoloških dostignuća, umjetna inteligencija bi trebala funkcionirati poput ljudskog bića s određenim stupnjem autonomnih sposobnosti, što omogućuje umjetnoj inteligenciji donošenje informiranih odluka na temelju skeniranog okruženja. Interakcije s umjetnom inteligencijom također su dizajnirane da se doimaju kao interakcije između ljudi jer su agenti umjetne inteligencije razvijeni da budu intimniji, bliskiji, utjelovljeni i sličniji ljudima (Westerman i sur., 2020).

2.2. Povijest i dosadašnji napredak umjetne inteligencije

Prema Russellu (2020) smatra se da je početak umjetne inteligencije (AI) kao definiranog područja znanosti bio 1956. kada su matematičari John McCarthy i Marvin Minsky surađivali s Claudeom Shannonom i Nathanielom Rochesterom na organizaciji ljetnog programa na Dartmouth Collegeu. Cilj je bio ambiciozan: istražiti mogućnost opisivanja svih aspekata učenja i inteligencije na dovoljno precizan način da ga strojevi mogu replicirati. Nastojali su strojevima omogućiti korištenje jezika, razvijanje apstrakcija i koncepata, rješavanje problema tradicionalno rezerviranih za ljude, pa čak i poboljšanje vlastitih sposobnosti. (ibid.) Unatoč njihovom početnom cilju za napredak unutar jednog ljeta, izazovi su se pokazali složenijim i

dugotrajnijim, protežući se daleko izvan tog vremenskog okvira. (ibid.) Umjetna inteligencija je postigla značajne prekretnice u svojim ranim fazama. Ti su uspjesi uključivali pionirski algoritam Alana Robinsona sa širokom primjenom. (ibid.) Međutim, složenost zadataka umjetne inteligencije i nijanse ljudske inteligencije značile su da se razvoj polja proširio daleko iznad početnih očekivanja. Razvoj umjetne inteligencije od početaka 1956. do tekućih izazova i uspjeha pokazuje i ambicioznu prirodu ciljeva tog područja i zamršenost uključenu u stvaranje strojeva koji mogu oponašati ljudske kognitivne sposobnosti. Temelj današnjeg napretka u području umjetne inteligencije postavljen je tijekom razdoblja poznatog kao "zima umjetne inteligencije". U ovom razdoblju pojavili su se rani napori u razvoju opsežnih probabilističkih sustava zaključivanja i temelja onoga što sada prepoznajemo kao duboko učenje. (ibid.) Od 2011. godine, tehnike dubinskog učenja donijele su značajne pomake u ključnim izazovima umjetne inteligencije, kao što su prepoznavanje govora, vizualno prepoznavanje objekata i strojno prevođenje. To su povijesno bili među najzahtjevnijim problemima na tom području. Strojevi su dosegli točku u kojoj su na razini ili čak premašuju ljudske sposobnosti u tim domenama (ibid.).



Slika 1. Primjena umjetne inteligencije (Izvor: Bird Academy, 2023)

Prvi pokušaji razvoja umjetne inteligencije pokazali su zamršenost područja i potrebu za uvođenjem i savladavanjem drugih disciplina kako bi se ona postigla. (Russell, 2020) Iz slike je vidljivo kako područje umjetne inteligencije obuhvaća područja:

• strojnog učenja,

• obrade prirodnog jezika,

• računalni vid,

• obradu govora,

• ekspertne sustave,

• planiranje i

• robotiku.

2.3. Pitanje strojne heuristike

Heuristika se može smatrati tehnikom pojednostavljenja. Heuristike su pristupi koji se često koriste kako bi se došlo do rješenja koje možda nije besprijekorno, ali je dovoljno točno za zahtjeve zadatka koji neposredno treba razriješiti. U području računalstva, heuristika se pokazala osobito vrijednom kada je postizanje idealnog rješenja problema neizvedivo zbog čimbenika kao što su mala brzina ili ograničenja procesorske snage.

Iako umjetna inteligencija ima značajke slične ljudskim i strojnim, fokus je često na karakteristikama umjetne inteligencije sličnim strojnim, zanemarujući njenu autonomnu sposobnost. Stoga, nakon što se izvor interakcije identificira kao umjetna inteligencija, ljudi će teže procijeniti interakcije na temelju karakteristika stroja koje im padnu na pamet (Sundar i Kim 2019).

Ovaj psihološki mehanizam ima mnogo paralela s objašnjenjem socijalne kognicije, koja opisuje ljude kao kognitivne „škrtice“. Umjesto prolaska kroz napornu procjenu, ljudi se oslanjaju na mentalnu heuristiku koju pokreću kontekstualni znakovi (Jones-Jang i Jin Park, 2023).

Postoji nekoliko razloga zašto se korisnici više oslanjaju na mentalne prečace nego na naporne procjene očekivane izvedbe umjetne inteligencije. Prvo, s gledišta korisnika, umjetna inteligencija je crna kutija (Shin i Park, 2019). Iako pokret koji stoji iza „razložive“ umjetne

inteligencije zagovara i potiče potrebu za transparentnošću u stvaranju i primjeni umjetne inteligencije, korisnicima je gotovo nemoguće znati kako je umjetna inteligencija programirana iza sučelja. Drugo, unatoč nedavnom porastu aplikacija umjetne inteligencije, mnogi korisnici još uvijek imaju malo izravnih interakcija s agentima umjetne inteligencije. Bez prethodnih iskustava i kognitivnih procesa korisnika, njihovo razumijevanje umjetne inteligencije bit će vrlo subjektivno i ograničeno (Just i sur., 2018.).

Takav nedostatak transparentnosti i prethodna izravna iskustva dovode do ideje da su početne percepcije ljudi o umjetnoj inteligenciji često oblikovane kroz medijske prezentacije umjetne inteligencije (Banks, 2020.). Početna percepcija ljudi o strojevima ili umjetnoj inteligenciji obično uključuje uvjerenje da umjetna inteligencija funkcionira na unaprijed programiran, objektivan i dosljedan način (Sundar i Kim, 2019).

Literatura o strojnoj heuristici sugerira da proces prosuđivanja slijedi logiku ako-onda-dakle. Na primjer, ako je izvorni znak prepoznat kao umjetna inteligencija, tada se pokreće heuristika stroja (programirana i dosljedna izvedba), čime se utječe na prosudbu korisnika o iskustvu umjetne inteligencije. Vrijedno je napomenuti da se strojna heuristika ne pojavljuje uvijek automatski ili ravnomjerno (Jones-Jang i Jin Park, 2023).

S obzirom na različite pretpostavke o, i različito razumijevanje umjetne inteligencije kojima ljudi raspolažu, da se tu obično radi o krajnostima pristranosti prema strojevima ili pak averziji prema njima, kao i da jedno negativno iskustvo s umjetnom inteligencijom može izazvati potpuno nepovjerenje korisnika u njenu budućnost, velika je odgovornost kod razvoja umjetne inteligencije osigurati transparentnost. Tako će se korisniku omogućiti određeni uvid i kontrola nad procesom strojnog donošenja odluka i varijablama koje stoje iza njega i utječu na rezultate. Ukoliko to nije opcija, korisniku je barem potrebno ponuditi ispravne znakove koji pokreću heuristiku stroja. Samim time omogućuje se i točnija korisnikova procjena pouzdanosti umjetne inteligencije i evaluaciju zaključaka koji iz nje proizlaze. Iz većeg razumijevanja i kontrole koju korisnik nad umjetnom inteligencijom ima, kao i ispravnim poticanjem heuristike stroja kod korisnika, postiže se i veće povjerenje.

Stoga, istraživači tvrde da bi ispravni znakovi trebali biti vizualno dostupni za pokretanje heuristike stroja. Štoviše, isti znak može potaknuti različite heuristike ovisno o kontekstu i korisnicima. Stoga je, prema ovoj literaturi, preporučena praksa u strojnom heurističkom istraživanju da se znakovi, strojna heuristika i prosudbe izravno mjere i tretiraju kao različite varijable (Jones- Jang i Jin Park, 2023).

3. PREDRASUDE UGRAĐENE U ALGORITME I AUTOMATIZIRANE PROCESE

U nastavku rada obrađuju se predrasude ugrađene u algoritmima i automatiziranim procesima gdje će se prikazati ujedno i konkretni primjeri gdje je navedeno uviđeno. Naime, predrasude ugrađene unutar algoritama i automatiziranih procesa uzrokuju brojne probleme. Navedeni problemi mogu biti različite prirode pa je stoga nužno da se predrasude reguliraju, odnosno da se navedene u što većem broju slučajeva izbjegnu.

3.1. Problematika pristranosti automatizacije i algoritamska averzija

Heuristički vođeni mehanizmi evaluacije dovode do dviju suprotnih percepcija umjetne inteligencije: pristranosti automatizacije i algoritamske averzije. Pristranost automatizacije javlja se kada korisnici precjenjuju dosljednu izvedbu i preciznost umjetne inteligencije, stvarajući idealistički okvir za mogućnosti umjetne inteligencije. Empirijski dokazi pokazuju da su ljudi skloniji prihvatiti savjete koje je generirala umjetna inteligencija nego ljudske savjete, posebno kada je riječ o predviđanjima i glazbenim preporukama (Logg i sur., 2018),

Nadalje, priče generirane umjetnom inteligencijom i poruke za provjeru činjenica više potiču korisnike da ublaže vlastite pristranosti i stavove vezane uz temu, te prihvate uzeti u obzir priču koja nudi stajalište različito od njihovog i naknadno ocjenjuju poruke koje su sastavili inteligentni sustavi od onih koje su izradili ljudi. U primjeni, informacije koje generira ili obrađuje umjetna inteligencija nisu imune na pristranosti zbog podataka koji pokreću odluke umjetne inteligencije. Unatoč tome, velik broj korisnika ostaje nesvjestan implicitnih pristranosti izazvanih umjetnom inteligencijom i doživljava umjetnu inteligenciju kao nepristranu (Jones-Jang i Jin Park, 2023.).

Za razliku od pristranosti automatizacije, gdje korisnici gaje optimistična predviđanja unaprijed programiranih performansi umjetne inteligencije, algoritamska averzija se očituje kada signali izvedeni iz stroja izazivaju nepovoljne odgovore prema umjetnoj inteligenciji. Prema Jonesu-Jangu i Jin Parku (2023), Ranija istraživanja (e.g. Alvarado-Valencia i Barrero, 2014;Bucher, 2017;Dietvorst i sur., 2015) razotkrivaju primjere algoritamske averzije ističući sklonost

korisnika da daju prednost ljudskim prosudbama nad umjetnom inteligencijom, čak i kada umjetna inteligencija nadmašuje ljude.

Često citirani razlozi iza algoritamske averzije poput odstupanja od idealne sheme izvedbe umjetne inteligencije, percipirane nesposobnosti umjetne inteligencije da uključi kontekstualne elemente i etičkih problema povezanih s oslanjanjem na strojeve za posljedične odluke. Značajna razlika u percepciji korisnika o pristranosti umjetne inteligencije nasuprot ljudskih agenata leži u pretjerano visokim očekivanjima korisnika o dosljednoj izvedbi umjetne inteligencije. To dovodi do povećanog razočaranja kada rezultati umjetne inteligencije budu loši. Posljedično, teoretizira se da pozitivne i negativne procjene umjetne inteligencije, odnosno pristranost prema automatizaciji i algoritamska averzija, čine sekvencijalne, a ne kontradiktorne procese u korištenju umjetne inteligencije. Korisnici se često pridržavaju idealističke sheme za AI, očekujući svaki put besprijekornu izvedbu. Posljedično, odstupanje od ovih visokih očekivanja znatno umanjuje kasniju sklonost korisnika AI-ju. Kvarovi do kojih dolazi zbog automatiziranih sustava često potiču korisnike da izbjegavaju njihovu upotrebu zajedno s umjetnom inteligencijom. Suprotno tome, budući da korisnici imaju realističnija očekivanja od ljudskih agenata i priznaju ljudsku pogrešivost u donošenju odluka, skloni su pokazati veću toleranciju na ljudske pogreške u usporedbi s onima koje čine sustavi umjetne inteligencije (Jones-Jang i Jin Park, 2023).

3.1.1. Procjena pravednosti - algoritamska averzija

Značenje pravednih procjena umjetne inteligencije naglašeno je u nedavnim studijama (Helberger i sur., 2018). Pravednost u umjetnoj inteligenciji znači percipirano očekivanje da umjetna inteligencija ne bi trebala donositi nepravedne, loše ili diskriminirajuće odluke. Studije (Shin i Park, 2019) su otkrile da se u utjecanju na upotrebu umjetne inteligencije i oslanjanje na njezine odluke, pravednost pojavila kao ključno normativno očekivanje, a posebno zato što se korisnici suočavaju s neizvjesnostima „crne kutije“, odnosno odlučivanja temeljenih na strojevima. Drugim riječima, za razliku od odluka koje donose ljudi, očekivanje pravednosti može biti još kritičnije u pogledu normative, tj. kazneno, kada se korisnici moraju osloniti na strojeve za važne odluke na koje ne mogu utjecati, ali ih ipak pokušavaju razumjeti (Jones-Jang i Jin Park, 2023).

Uostalom, korisnička iskustva s agentima umjetne inteligencije mogu se temeljno temeljiti na njihovim subjektivnim procjenama i očekivanjima, a pravednost može biti istaknuti atribut u ovom procesu, posebno za odluke stroja koje im nisu u korist. Potrebno je imati na umu da se ova procjena pravednosti razlikuje od procjene agenata umjetne inteligencije za njihovu točnost. Umjesto toga, zabrinutost je u slučaju „loših“ odluka umjetne inteligencije i automatiziranih procesa koje će vjerojatno dovesti do zahtjeva za boljim objašnjenjima. Ovaj zahtjev za pravednošću može biti funkcionalni barometar u potrazi korisnika za tim kako percipirati nepovoljne odluke koje nisu donijeli ljudi (Jones-Jang i Jin Park, 2023).

Istraživanje koje su proveli Jones-Jang i Jin Park bilo je u segmentu procjene pravednosti za agenta umjetne inteligencije (računalo), različito od ljudskog agenta, a ključ je identificirati kako pojmovi pristranosti automatizacije i algoritamske averzije zajedno objašnjavaju mehanizam procjene. Konkretno, istraživanjem je predviđeno da će umjetna inteligencija vjerojatnije raditi dosljednije od ljudi (automatizirana pristranost) te kako ljudi smatraju da je vjerojatnije da će umjetna inteligencija dati negativnu ocjenu pravednosti pogrešnih odluka umjetne inteligencije (algoritamska averzija). U tom segmentu tumači se kako će ljudi pokazati negativnije procjene pravednosti agenta umjetne inteligencije, u usporedbi s ljudskim agentom (algoritamska averzija), jer su njihova početna očekivanja o dosljednoj izvedbi umjetne inteligencije (automatizirana pristranost) prekršena (Jones-Jang i Jin Park, 2023).

3.1.2. Percipirana mogućnost kontrole

Prevladavajuća perspektiva u vezi s umjetnom inteligencijom koja može oblikovati procjenu njezine izvedbe odnosi se na njezinu percipiranu mogućnost kontrole. Dok su ranija istraživanja o strojnoj heuristici uvelike zanemarivala koncept percipirane upravljivosti u procjeni umjetne inteligencije, suvremene studije razjašnjavaju da početne percepcije korisnika o tome u kojoj mjeri automatizirani procesi i umjetna inteligencija mogu utjecati na ishode predstavlja jednu od ključnih heuristika stroja uz automatizacijsku pristranost i algoritamsku averziju, utječući na kasnije prosudbe o umjetnoj inteligenciji, uključujući procjene pravednosti (Jones-Jang i Jin Park, 2023).

Okvir atribucijske teorije, kako ju je postavio Weiner (2006), istražuje kauzalnost i percepciju odgovornosti, točnije kako pojedinci pripisuju uzroke posljedicama i, ovisno o tome jesu li čimbenici koji su izazvali akciju i utjecali na ishod vanjski ili unutarnji, odnosno razini kontrole

nad čimbenicima koji mu prethode i ishodom, kako objašnjavaju i opravdavaju određene postupke i njihove rezultate. On pretpostavlja da pozitivne ili negativne procjene pojedinaca o određenim entitetima ovise o tome mogu li ti entiteti utjecati na posljedice. Kada se smatra da je negativan ishod izvan kontrole entiteta, krivnja se smanjuje. Suprotno tome, ako se isti negativni ishod percipira kao dio ovlasti entiteta, krivnja se pojačava. Na primjer, pripisivanje nepovoljnih ishoda čimbenicima koji su pod kontrolom entiteta, poput neadekvatnog truda ili pogrešaka, daje strože ocjene, dok pripisivanje ishoda nekontroliranim čimbenicima, kao što je sreća, izaziva suosjećanje i pozitivne procjene (Weiner, 2006).

Prenoseći ovu logiku na područje umjetne inteligencije, korisničke procjene umjetne inteligencije ovise o njihovoj percepciji utjecaja umjetne inteligencije na ishode. U literaturi se naglašava da se rijetko smatra da umjetna inteligencija i roboti posjeduju kontrolu na ljudskoj razini. Korisnici često pripisuju mjesto kontrole vanjskim izvorima, usmjeravajući krivnju prema najbližim agentima kao što su organizacije koje usvajaju AI, programeri softvera koji ga programiraju ili korisnici koji s njim komuniciraju (Shank i DeSanti, 2018).

Alternativno stajalište koje nude van der Woerdt i Haselager (2016) tvrdi da AI posjeduje određeni stupanj kontrole nad ishodima zbog svoje očekivane neovisne i autonomne prirode u odnosu na druge strojeve. Međutim, ova se usporedba odnosi na AI agente u odnosu na druge strojeve, a ne na kontrast između AI i ljudskih agenata. Stoga, bez obzira na percepciju da umjetna inteligencija utjelovljuje različite stupnjeve inteligencije i autonomije, prevladava uvjerenje da umjetnoj inteligenciji nedostaje potpuna neovisnost i postoji skepticizam u pogledu njezine sposobnosti da zamijeni ljudsku inteligenciju i snosi odgovornost (Sundar i Kim 2019). Stoga, predviđa se da će pojedinci, vođeni heuristikom da umjetna inteligencija ima manju kontrolu nad ishodima, manje kriviti AI za negativne posljedice u usporedbi s ljudskim akterima (Jones-Jang i Jin Park, 2023).

3.2. Automatizacija i algoritamska averzija kao problem

Privatni i javni sektor postupno prihvaćaju sustave umjetne inteligencije i algoritme strojnog učenja kako bi pojednostavili osnovne i zamršene postupke donošenja odluka. Opsežna konverzija podataka u digitalni format i kasnije korištenje novih tehnologija uzrokuju značajne preokrete u brojnim gospodarskim sektorima, uključujući prijevoz, maloprodaju, oglašavanje, energiju i razna druga područja. Nadalje, utjecaj umjetne inteligencije proteže se na područje

demokracije i upravljanja, gdje se implementiraju računalni sustavi kako bi se povećala preciznost i potaknula nepristranost unutar vladinih operacija.

Prisutnost opsežnih skupova podataka pojednostavila je proces izvlačenja svježih uvida pomoću računala. Posljedično, algoritmi, koji obuhvaćaju skupove sustavnih uputa koje slijede računala za izvršavanje zadataka (Blass i Gurevich, 2006), razvili su se u sofisticirane i široko rasprostranjene alate za automatizirano donošenje odluka. Iako algoritmi imaju različitu primjenu, ovaj se rad usredotočuje na računalne modele koji izvlače uvide iz podataka o pojedincima, uključujući njihove identitete, demografske značajke, preferencije i predvidiva ponašanja, kao i povezane entitete.

Prema Chodosh (2018), u eri koja je prethodila algoritmima, i pojedinci i organizacije bili su odgovorni za odluke vezane uz zapošljavanje, oglašavanje, kazne i zajmove. Te su odluke često bile regulirane pravnim propisima na saveznoj, državnoj i lokalnoj razini koji su nalagali pravednost, transparentnost i ravnopravnost u procesima donošenja odluka. U trenutnom okruženju, neke od tih odluka u potpunosti se provode ili su pod utjecajem strojeva, koji se mogu pohvaliti iznimnom učinkovitošću zbog svoje veličine i statističke preciznosti. Algoritmi koriste opsežne makro i mikro podatke kako bi utjecali na odluke koje obuhvaćaju široku lepezu zadataka, od ponude filmskih preporuka do pomoći bankama u procjeni kreditne sposobnosti pojedinaca. Unutar strojnog učenja, algoritmi ovise o različitim skupovima podataka ili podacima o obuci kako bi se izveli točni ishodi. Putem ovih podataka o obuci oni stječu uvide i konstruiraju modele primjenjive na druge pojedince ili subjekte, omogućujući predviđanja o odgovarajućim ishodima za njih (Chodosh, 2018.).

Prije nego što se uključi u algoritamsko odlučivanje, čovjek koji donosi odluke rijetko se susreće s algoritmom bez očekivanja odnosno pretpostavki. Ljudi često imaju unaprijed formirana predviđanja u vezi mogućnosti, funkcija i prikladnosti algoritama. Ta predviđanja mogu proizaći iz neposrednog poznavanja algoritamskih alata, iskustva specifičnog za domenu ili neizravnog znanja prikupljenog od kolega i medija. Ishod ovih postojećih predviđanja dovodi do paradigme u kojoj ljudi koji donose odluke tumače i odgovaraju na savjete koje generiraju algoritmi različito u usporedbi sa savjetima ljudi, čak i kada je sadržaj savjeta isti. U literaturi su vidljivi različiti mehanizmi koji leže u pozadini ove razlike u odgovoru. To uključuje sklonost ljudi da uspostave društvene ili parasocijalne veze s izvorima savjeta, trajno uvjerenje da su ljudske pogreške nasumične i da ih je moguće ispraviti, dok su algoritamske pogreške sustavne, povjerenje stručnjaka u području koje dovodi do nedovoljne upotrebe naizgled nepotrebnih algoritamskih pomagala i slučajeve u kojima nedostatak obučavanja onemogućuje

ljudskom korisniku optimalnu upotrebu algoritamskog pomagala. U biti, očekivanja koja ljudski korisnik donosi u interakciju između ljudi i algoritama značajno utječu na to kako on koristi algoritam (Burton i sur., 2020).

Da bi osoba koja donosi odluke mogla djelovati prema procjeni algoritma, mora imati osjećaj kontrole i dovoljno povjerenja. Ovaj osjećaj kontrole može proizaći iz istinskog razumijevanja izvedbe algoritma, ali također može proizaći iz prilagodbi algoritamskog procesa donošenja odluka koje ne utječu bitno na stvarno funkcioniranje algoritma (kao što je mijenjanje sučelja za prezentaciju informacija uz zadržavanje pristupa analizi informacija algoritma). Povjerenje u pomoć pri odlučivanju, kalibrirano na temelju čimbenika kao što su predvidljivost, pouzdanost, tehnička kompetencija, uzajamnost i moralnost, dovodi do uvjerenja da pomoć služi dobronamjernoj svrsi, a ne prijevari ili kontroli. U kontekstu algoritamskih pomagala za donošenje odluka, naglašava se važnost pružanja stvarne ili percipirane kontrole odlučivanja ljudskim korisnicima kako bi se odgovorilo na njihove psihološke potrebe i osobni interes. Povjerenje u algoritam brzo erodira nakon opažanja pogrešaka, ali se ipak može odmah vratiti omogućavanjem ljudskom donositelju odluka da prilagodi prosudbu algoritma, čak i unutar određenih ograničenja. Time se naglašava manifestacija odbojnosti prema algoritmima u poboljšanim sustavima donošenja odluka koji ne uspijevaju odgovoriti na psihološke zahtjeve ljudskih korisnika za djelovanjem, autonomijom i kontrolom (Burton i sur., 2020).

4. UZROCI PREDRASUDA I PRISTRANOSTI UGRAĐENIH U ALGORITME I AUTOMATIZIRANE PROCESE

Kako se tehnologija ubrzano integrira u različite aspekte svakodnevice, automatizirani procesi i algoritmi igraju sve utjecajniju ulogu u donošenju odluka. Dok ovi sustavi obećavaju nepristranost i učinkovitost, oni također iznose na vidjelo nešto zabrinjavajuće – potencijal za ugrađene predrasude. Unatoč svojoj reputaciji objektivnosti, automatizirani procesi i algoritmi mogu nenamjerno ovjekovječiti pristranosti koje su prisutne u podacima na kojima se obučavaju ili u dizajnerskim odlukama koje su napravili njihovi kreatori. Ovo pitanje postavlja daljnja pitanja o pravednosti, jednakosti i etičkim implikacijama oslanjanja na te sustave za donošenje kritičnih odluka (Silberg i Manyika, 2019). U ovom dijelu teksta istražuju se uzroci predrasuda ugrađenih u automatizirane procese i algoritme, kako se predrasude mogu pojaviti i ovjekovječiti, utjecaj tih predrasuda na marginalizirane skupine i imperativ rješavanja ovog izazova kako bi se osigurali pravedni i nepristrani ishodi.

4.1. Strategije detekcije predrasuda i pristranosti ugrađenih u algoritme i automatizirane procese

Od ključne je važnosti rješavanje averzije prema algoritmima povećanjem algoritamske pismenosti među ljudima koji donose odluke. To uključuje obuku pojedinaca ne samo u njihovim profesionalnim domenama, već i u interakciji s algoritamskim alatima, razumijevanju statističkih rezultata i uvažavanju pomoći pri odlučivanju. Algoritamska pismenost obuhvaća razumijevanje ključnih statističkih koncepata kao što su pogreška i nesigurnost. Na primjer, stopa točnosti od 80% u donošenju odluka često je eksplicitna, čime je jasna mogućnost netočnosti od 20%. Nasuprot tome, stope pogrešaka korisnika obično se ne spominju, zbog čega podcjenjuju izvedbu algoritma. Međutim, oslanjanje isključivo na algoritamsku pismenost prebacuje odgovornost na korisnike i možda neće u potpunosti spriječiti averziju prema algoritmu jer zanemaruje šire faktore donošenja odluka i dizajn alata. Štoviše, učinkovitost takvog opismenjavanja može biti ograničena jer studije o averziji prema algoritmima često uključuju sudionike s postojećim algoritamskim znanjem. Stoga, samo oslanjanje na algoritamsku pismenost možda neće biti sveobuhvatno rješenje za rješavanje averzije prema algoritmu (Burton i sur., 2020).

Sve veći naglasak stavljen je i na koncept donošenja odluka u petlji, gdje ljudski korisnici surađuju s algoritmima kako bi poboljšali donošenje odluka. Ovaj pristup omogućuje korisnicima da interveniraju, daju ulazne podatke i utječu na konačnu odluku, a može imati različite oblike kao što su interaktivni sustavi podrške, sustavi ljudske automatizacije ili angažirani sustavi. Ljudi su relativno ravnodušni u mjeri u kojoj mogu modificirati nesavršene prognoze algoritma, sve dok mogu dati svoj doprinos i sudjelovati u konačnoj odluci. Ovo sugerira da čak i iluzija autonomije može ublažiti odbojnost prema algoritmu. Kako bi se to riješilo, prošireni sustavi odlučivanja trebali bi uključivati bihevioralne znakove ili čimbenike vjerodostojnosti koji su manje važni za izvedbu donošenja odluka, ali ključni za prevladavanje averzije prema algoritmu. Ovaj pristup kombinira načela estetskog dizajna i funkcionalnosti. Međutim, uključivanje takvih značajki može dovesti do sporijeg izvlačenja odluka, što ovo rješenje čini prikladnim za domene u kojima je izvedivo dovoljno vremena za suradnju s algoritamskim pomagalima (Burton i sur., 2020).

Iako postoje nedosljednosti u literaturi o poticanju korištenja algoritama, razumno je pretpostaviti da motiviranje ljudi koji donose odluke da prihvate algoritamsku prosudbu zahtijeva svjesno oblikovanje konteksta odluke. To uključuje rješavanje averzije prema algoritmu kao projekta promjene ponašanja, gdje ukorijenjene organizacijske rutine i društvene norme predstavljaju značajne prepreke. Predloženi su različiti prijedlozi u tom smjeru: prednosti algoritma za uokvirivanje u terminima koji se mogu usporediti, manipuliranje percipiranim društvenim konsenzusom, a Fisher (2008), lokalizirani sustavi nagrađivanja prilagođeni određenim ulogama u donošenju odluka u organizacijama. Iako svaki pristup pokazuje potencijal, najučinkovitija strategija poticanja vjerojatno će ovisiti o kontekstu. Stoga uspješna implementacija algoritamske odluke može zahtijevati dizajn ponašanja specifičan za situaciju (Burton i sur., 2020).

4.2. Primjeri predrasuda i pristranosti ugrađenih u algoritme i automatizirane procese

Algoritamska pristranost može se pojaviti na nekoliko načina i s različitim stupnjevima posljedica. U nastavku će se navesti neki od primjera. Kao prvi primjer navodi se pristranost u online alatima namijenjenim za zapošljavanje.

Amazon, div e-trgovine s globalnom radnom snagom koju sačinjava 60% muškaraca, a 74% vodećih pozicija čine muškarci , nedavno je obustavio korištenje algoritma zapošljavanja zbog otkrivanja rodne pristranosti. Razvoj algoritma temeljio se na podacima dobivenim iz

životopisa poslanih Amazonu tijekom desetljeća, prvenstveno od muških kandidata. Umjesto da se usredotoči na relevantne skupove vještina, algoritam je uvježban za prepoznavanje uzoraka riječi unutar životopisa. Ti su obrasci zatim uspoređeni sa sastavom pretežno muškog inženjerskog odjela kako bi se procijenila prikladnost kandidata. Posljedično, softver umjetne inteligencije pokazao je oblik pristranosti negativno utječući na životopise koji u svom sadržaju sadrže izraz "žensko", što je rezultiralo nepravednim postupanjem prema kandidatkinjama (Vincent, 2018).

Drugi primjer je algoritamska pristranost u algoritmu kaznenog pravosuđa u Americi. Otkriveno je da algoritam COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) koji koriste suci za predviđanje hoće li optuženici biti zadržani u pritvoru ili pušteni uz jamčevinu prije suđenja, pokazuje predrasudu prema Afroamerikancima. Ovaj algoritam generira ocjenu rizika koja ukazuje na vjerojatnost da će okrivljenik počiniti budući zločin korištenjem niza izvora podataka, uključujući evidenciju uhićenja i demografiju okrivljenika. Istragom je utvrđeno da će, u usporedbi s pojedincima bijele rase s jednakom vjerojatnošću ponavljanja kaznenog djela, Afroamerikanci vjerojatnije dobiti višu procjenu rizika od algoritma, što je posljedično dovelo do produljenih razdoblja pritvora prije suđenja (Corbett-Davies i sur., 2017).

Pristranost je također uočena u tehnologiji prepoznavanja lica. Joy Buolamwini, istraživač s MIT-a, otkrila je da algoritmi koji se koriste u tri komercijalno dostupna softverska sustava za prepoznavanje lica pokazuju nedostatke u prepoznavanju osoba s tamnijim tonovima kože. Općenito, procjenjuje se da se većina skupova podataka za obuku prepoznavanja lica sastoji od preko 75% muškaraca i preko 80% pojedinaca bijele rase. Kada je subjekt fotografije bijel, softver ga je ispravno identificirao kao muškarca u 99% slučajeva. Buolamwinijevo istraživanje otkriva da su kombinirane stope pogreške proizvoda za tri sustava bile manje od 1% sveukupno, ali te su stope eskalirale na više od 20% za jedan proizvod i 34% za druga dva proizvoda kada se pokušalo identificirati žene tamnije puti kao žene. Slijedeći Buolamwinijeva otkrića u vezi s analizom lica, i IBM i Microsoft obvezali su se poboljšati točnost svog softvera u prepoznavanju lica s tamnijim tonovima kože (Hardesty, 2018).

Ovi primjeri pristranosti, iako nisu iscrpni, naglašavaju da ta pitanja nisu samo teoretska, već konkretna u praksi. Oni ilustriraju mehanizme kroz koje nastaju takvi ishodi, često bez ikakve loše namjere kreatora ili operatera algoritma. Prepoznavanje potencijala i podrijetla pristranosti predstavlja početni korak u rješavanju takvih problema kroz strategije ublažavanja.

5. ZAKLJUČAK

Brzi napredak tehnologije duboko ju je integrirao u različite aspekte modernog svakodnevnog života, kako u privatnoj tako i poslovnoj sferi. Sustavi umjetne inteligencije sve se više oslanjaju na automatizaciju brojnih odluka i procesa. Strojevi se percipiraju kao precizni, analitički i nepristrani, zbog čega su odluke izvedene iz algoritama strojnog učenja poželjnije zbog njihove percipirane objektivnosti. Međutim, ovaj rad otkriva da ni umjetna inteligencija nije izuzeta od problema i nepravde. Kako ovi sustavi nalaze sve više primjena, njihove nesavršenosti i pristranosti mogu značajno utjecati na ljudska prava i slobode, potencijalno dovodeći do zlouporabe.

Taj kontekst naglašava hitnu potrebu za propisima, standardizacijom i sveobuhvatnim pregledom algoritama koji imaju značajan utjecaj na ljudske živote. Kroz ovaj rad iznose se činjenice koje ukazuju na to da sustavi umjetne inteligencije, ili algoritmi, mogu gajiti predrasude. Ovaj primjer rezultira "pristranošću", koju općenito definiramo kao sustavno nepovoljne ishode za određene skupine bez opravdanih razlika među tim skupinama. Pristrani algoritmi mogu proizaći iz neadekvatnih ili nereprezentativnih podataka o obuci i oslanjanja na pogrešne informacije koje odražavaju povijesne nejednakosti. Ako se ne riješe, takve predrasude mogu dovesti do odluka koje imaju različit utjecaj na određene skupine, čak i kada razvojni programeri nemaju namjeru diskriminacije. Stoga postaje ključno primijeniti stroge procese kako bi se predrasude i pristranosti svele na najmanju moguću mjeru.

LITERATURA

1. Banks, J. (2020). Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI, 7*, 62.
2. Blass, A., & Gurevich, Y. (2006). Algorithms: A quest for absolute definitions. In: Olszewski, A., Wolenski, J. & Janusz, R. (Eds.) (2006). *Church's Thesis After 70 Years*. Berlin, Boston: De Gruyter. 24–57.
3. Burton J. W., Stein M. K., & Jensen T. B. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239.
4. Chodosh, S. (2018). Courts use algorithms to help determine sentencing, but random people get the same results. *Popular Science*. https://www.popsci.com/recidivism-algorithm-random-bias [05. veljače 2023.].
5. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In: Matwin, S., Yu, S., & Farooq, F. (Eds.) (2017). *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery. 797–806.
6. Fisher, C. D. (2008). Why don't they learn? *Industrial and Organizational Psychology*, *1*(3), 364–366.
7. Hardesty, L. (2018). Study finds gender and skin-type bias in commercial artificial-intelligence systems. *MIT News Office.* https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 [26. rujna 2023.].
8. Helberger N., Karppinen K., & D'Acunto L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society, 21*(2), 191–207.
9. Jones-Jang, S. M., & Jin Park, Y. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication, 28*(1), 1–13.
10. Just N., Latzer M., & Warf B. (2018). Algorithmic selection on the Internet. In: Warf B. (Eds.), *The SAGE encyclopedia of the internet.*
11. Logg, J. M., Minson, J. A., & Moore, D. A. (2018). *Algorithm appreciation: People prefer algorithmic to human judgment.* Harvard Business School.

12. Russell, S. J. (2020). Human compatible. In: *Artificial intelligence and the problem of control*, str. 8–10. New York: Penguin Books.
13. Silberg, J., & Manyika, J. (2019). *Notes from the AI frontier: Tackling bias in AI (and in humans)*. McKinsey Global Institute.
14. Shabbir, J., & Anwer, T. (2018). *Artificial Intelligence and its Role in Near Future*. https://arxiv.org/abs/1804.01396 [1. veljače 2023.].
15. Shank D. B., & DeSanti A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in human behavior, 86*, 401–411.
16. Shin D., & Park Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277–284.
17. Sundar S. S., & Kim J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In: Brewster, S. A., Fitzpatrick, G., Cox, A. L., Kostakos, V. *Proceedings of the 2019 Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery.1–9.
18. van der Woerdt, S., & Haselager, P. (2017). Lack of effort or lack of ability? robot failures and human perception of agency and responsibility. *Communications in Computer and Information Science*, 155–168.
19. Vincent, J. (2018). Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women. *The Verge.* https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report [31. siječnja 2023.].
20. Wamba-Taguimdje, S.-L., Fosso Wamba, S., Kala Kamdjoug, J. R., & Tchatchouang Wanko, C. E. (2020.). Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. *Business Process Management Journal, 26*(7): 1893–1924.
21. Weiner B. (2006). *Social motivation, justice, and the moral emotions: An attributional approach.* New Jersey: Lawrence Erlbaum Associates.
22. Westerman D., Edwards A. P., Edwards C., Luo Z., & Spence P. R. (2020). I-it, i-thou, i- robot: The perceived humanness of AI in human-machine communication. *Communicating Artificial Intelligence (AI)*, 25–37.

PRILOZI

Slika 1. Primjena umjetne inteligencije

Predrasude ugrađene u algoritme i automatizirane procese

SAŽETAK

Umjetna inteligencija (AI) je polje računalne znanosti koje se fokusira na stvaranje sustava sposobnih za obavljanje zadataka koji obično zahtijevaju ljudsku inteligenciju. Ovi zadaci uključuju rješavanje problema, donošenje odluka, razumijevanje prirodnog jezika i prepoznavanje uzoraka. Tehnologije umjetne inteligencije obuhvaćaju niz tehnika, poput strojnog učenja, neuronskih mreža i dubokog učenja, omogućujući računalima da uče iz podataka i poboljšavaju svoje performanse tijekom vremena. Jedna značajna primjena umjetne inteligencije je automatizacija procesa u raznim industrijama. To može dovesti do povećane učinkovitosti, smanjenja pogrešaka i uštede troškova. Industrije poput proizvodnje, logistike, korisničke službe i financija imale su koristi od automatizacije vođene umjetnom inteligencijom, budući da se sustavi umjetne inteligencije mogu brzo i precizno baviti ponavljajućim i dugotrajnim zadacima. Iako automatizacija koju pokreće AI nudi brojne prednosti, ona također donosi izazove. Premještanje radne snage zbog strojeva koji preuzimaju zadatke koje su prethodno obavljali ljudi predstavlja problem koji zahtijeva pažljivo upravljanje prijelazom radne snage. Štoviše, dizajn i implementacija automatiziranih procesa moraju uzeti u obzir etička razmatranja, potencijalne pristranosti i potrebu za ljudskim nadzorom kako bi se osiguralo odgovorno donošenje odluka. Automatizacija vođena umjetnom inteligencijom transformirala je industrije zamjenom manualnih zadataka učinkovitim i inteligentnim sustavima. Iako nudi značajne prednosti, zahtijeva pažljivo planiranje za rješavanje društvenih, etičkih i ekonomskih implikacija. Cilj ovog rada je ponuditi pregled i opis postojećih i nadolazećih tehnologija, predstaviti znanstvene radove i istraživanja, kao i projekte kojima je ovo u fokusu, te istaknuti njihovu viziju, ciljeve i dosadašnja postignuća.

Ključne riječi: umjetna inteligencija, tehnologija, strojno učenje, automatizirani procesi

Biases built into algorithms and automated processes

SUMMARY

Artificial Intelligence (AI) is a field of computer science that focuses on creating systems capable of performing tasks that typically require human intelligence. These tasks include problem-solving, decision-making, natural language understanding, and pattern recognition. AI technologies encompass a range of techniques, such as machine learning, neural networks, and deep learning, enabling computers to learn from data and improve their performance over time. One significant application of AI is in automating processes across various industries. This can lead to increased efficiency, reduced errors, and cost savings. Industries like manufacturing, logistics, customer service, and finance have benefited from AI-driven automation, as repetitive and time-consuming tasks can be handled swiftly and accurately by AI systems. While AI-powered automation offers numerous benefits, it also brings challenges. Workforce displacement due to machines taking over tasks previously done by humans is a concern, requiring careful management of workforce transitions. Moreover, the design and implementation of automated processes must consider ethical considerations, potential biases, and the need for human oversight to ensure responsible and accountable decision-making. AI-driven automation has transformed industries by replacing manual tasks with efficient and intelligent systems. While it offers substantial advantages, it requires careful planning to address social, ethical, and economic implications. The purpose of this paper is to give an overview and description of already existing and upcoming technologies, science articles and research papers, as well as projects with this topic in their focus, while emphasizing their vision, goals and achievements.

Keywords: artificial intelligence, technology, machine learning, automated processes