https://doi.org/10.18485/fdu_zr.2024.46.9

Katarina Šmakić[1]
Fakultet za diplomatiju i bezbednost, Beograd

316.32:004
316.776:004.738.5
COBISS.SR-ID 159226633

UTICAJ DIGITALNIH MEDIJSKIH SISTEMA NA DRUŠTVENU POLARIZACIJU

Apstrakt

Algoritmi za moderaciju postali su suštinski deo savremenih digitalnih medijskih sistema, omogućavajući automatsko filtriranje i uklanjanje neprikladnog sadržaja bez potrebe za ljudskom intervencijom. Ovaj rad istražuje implikacije takvog pristupa, s posebnim fokusom na njegov uticaj na polarizaciju unutar društva. Dok pružaju efikasnost u suzbijanju nasilnog i neprimerenog sadržaja, algoritmi u isto vreme funkcionišu na bazi unapred definisanih pravila koja mogu nenamerno filtrirati određene stavove ili informacije. Ovaj proces može pojačati stvaranje zatvorenih krugova informacija i uskih interesnih zajednica, pri čemu algoritmi favorizuju sadržaj koji se slaže s prethodnim interesovanjima korisnika, smanjujući njihovu izloženost suprotnim stavovima i dodatno doprinose polarizaciji. Zaključak ovog rada naglašava potrebu za razvijanjem sofisticiranih mehanizama prepoznavanja algoritamskih moderacija koje doprinose polarizaciji. Ključna preporuka je uvođenje pravnog okvira koji će obavezivati digitalne platforme na transparentnost algoritama, uključujući objavljivanje kriterijuma prema kojima se sadržaj filtrira i moderira, kao i podatke na kojima se treniraju. Pravne regulative bi trebalo da obuhvate mehanizme za nezavisnu reviziju i praćenje algoritamskih odluka, kao i uspostavljanje odgovornosti platformi za potencijalne negativne posledice na društveni dijalog i pluralizam mišljenja. Time bi se obezbedio balans između efikasnosti moderacije i zaštite osnovnih prava korisnika, kao što su sloboda izražavanja i pristup različitim stavovima.

Ključne reči

digitalni medijski sistemi, algoritamska moderacija, transparentnost, dijalog, sloboda izražavanja.

1 katarinasmakic@gmail.com; ORCID iD 0009-0003-3838-9525

Uticaj medijskih sistema na polarizaciju društva započinje još u 19. veku, kada su novine bile glavni izvor informacija i često se koristile kao politički alat za oblikovanje javnog mnjenja, promovisanje ideoloških stavova i podsticanje političkih sukoba između različitih društvenih grupa, čime su značajno doprinele stvaranju dubokih podela u društvu. Štampa nije bila neutralna; izdavači su otvoreno podržavali određene političke ideologije i interese elita, često manipulišući informacijama kako bi ojačali pozicije moćnih grupa, utičući na percepciju čitalaca i usmeravajući javne diskusije u pravcu koji je odgovarao njihovim sponzorima, što je dodatno pojačavalo političku polarizaciju i društvene tenzije (Bennett, Iyengar 2008: 711). Ova praksa je često dovodila do regionalnih podela, gde su različite novine prikazivale događaje iz perspektive suprotstavljenih strana, selektivno birajući informacije i interpretacije koje su odgovarale interesima određene političke grupacije ili lokalnih moćnika, stvarajući atmosferu nepoverenja među građanima i jačajući osećaj pripadnosti različitim ideološkim taborima, često rezultirajući oštrim sukobima i fragmentacijom društva na lokalnom i nacionalnom nivou (Herman, Chomsky 1988: 1).

U 20. veku, s razvojem radija i televizije, medijski uticaj postaje još širi. Iako su ovi mediji doneli homogenizaciju informacija na nacionalnom nivou, centralizovana kontrola nad sadržajem i ograničen broj kanala omogućili su vladi i korporacijama da oblikuju javno mnjenje prema vlastitim interesima. Medijski moćnici su često gurali agende koje favorizuju određene političke stavove ili ekonomske interese, stvarajući polarizovane reakcije u društvu, posebno u vreme političkih kriza ili ratova (Chomsky 2002: 12). Međutim, pravi skok u medijskoj polarizaciji dolazi u 21. veku s pojavom interneta i društvenih mreža, koji su omogućili brzu i široku distribuciju informacija, ali i dezinformacija, stvarajući stereo realnost gde se korisnici okupljaju oko sličnih stavova i potvrđuju sopstvene predrasude, dok su algoritmi dodatno pojačavali ovu dinamiku, favorizujući sadržaje koji izazivaju snažne emotivne reakcije, često ekstremizujući političke debate i produbljujući društvene podele, jer su ljudi sve više bili izloženi jednostranim informacijama i retko su se susretali s različitim perspektivama. Sam termin „društvena polarizacija" (Somer, McCoy 2018: 2) odnosi se na proces u kojem se društvo deli na različite, često suprotstavljene grupe koje imaju različite vrednosti, stavove ili interese. Ovaj fenomen može rezultirati jačanjem identitetskih razlika i stvaranjem rivalstva između grupa, što može dovesti do konflikata, otuđenja i napetosti unutar zajednice. Društvena polarizacija može biti povezana s različitim faktorima, uključujući ekonomske nejednakosti, političke razlike i društvene tenzije, a njen uticaj se može manifestovati kroz smanjenje so-

cijalne kohezije i povećanje ekstremizma. Bilo da se odnosi na socijalne ili ekonomske aspekte, ovakav pristup slabi srednju klasu i podstiče ekstremne razlike među grupama, jer polarizovana medijska scena često zanemaruje interese i potrebe šireg društva, dok istovremeno favorizuje ekstremne stavove i marginalizuje umerenost (Pariser 2011: 49). Srednja klasa, koja obično igra ključnu ulogu u održavanju društvene stabilnosti, sve je više pritisnuta između suprotstavljenih ideoloških struja, dok se ekonomske nejednakosti produbljuju, a socijalna mobilnost se sve više ograničava. Ovaj proces doprinosi slabljenju društvene kohezije i povećanju osećaja nepravde, čime se stvaraju plodni uslovi za politički radikalizam i dalje fragmentiranje društva. Njen kompleksan i višedimenzionalni karakter zahteva dublje razumevanje i preciznije metode merenja, dok razlikovanje polarizacije od nejednakosti naglašava potrebu za razvojem novih pristupa i modela analize.

U svetlu ovih izazova, važno je razvijati strategije koje doprinose smanjenju polarizacije i jačanju socijalne kohezije, stvarajući stabilnija i inkluzivnija društva (Maggino 2019), zato što se ovaj vid društvene fragmentacije često smatra „[...] merom ideološke ili socijalne distance između različitih grupa u društvu“ (McCoy, Jennifer et al. 2018). Miroljub Radojković navodi da je „[...] jedna od opasnosti svakako anonimnost kao izvora informacija u građanskom novinarstvu ... digitalni mediji, osim ako nisu onlajn verzija institucionalnih, mogu bez rizika da prenose dezinformacije i glasine“ (Radojković 2017). Anonimnost na internetu može podstaći ekstremnije izražavanje mišljenja, dok širenje dezinformacija dodatno pogoršava polarizaciju. Ove dinamike često dovode do smanjenja međusobnog poverenja, zaoštravaju sukobe i otežavaju dijalog između različitih grupa. Razlika između tradicionalnih i digitalnih medija u kontekstu polarizacije leži u načinu distribucije informacija, interakciji s publikom, brzini objavljivanja i mehanizmima filtriranja sadržaja. Tradicionalni mediji imaju centralizovanu kontrolu i pružaju jednostranu komunikaciju, gde mali broj urednika i vlasnika medija odlučuje koje će informacije biti prenete javnosti, dok digitalni mediji omogućavaju decentralizovano stvaranje i deljenje sadržaja, gde svaki korisnik može postati proizvođač informacija i direktno doprinositi javnom diskursu (Jenkins 2006). Ova promena omogućava veću participaciju u kreiranju sadržaja, ali i širenje širokog spektra informacija, mišljenja i narativa, bez istih uredničkih standarda ili odgovornosti, čime dolazi do veće raznolikosti glasova u javnoj sferi, ali i do lakšeg širenja nekontrolisanog protoka netačnih ili manipulativnih sadržaja.

Digitalni mediji koriste algoritme koji favorizuju emocionalno angažovan sadržaj (Vosoughi, Roy, Aral 2018), stvarajući posebna mesta koja jačaju podele, dok tradicionalni mediji obično nude uravnoteženije informacije i nisu toliko interaktivni. Brža distribucija informacija u digitalnom svetu često dovodi do toga da se vesti i sadržaji šire mnogo pre nego što su detaljno provereni ili analizirani, što može stvoriti konfuziju među korisnicima. Ovaj ubrzani tok informacija znači da ljudi imaju manje vremena da kritički razmotre ili provere tačnost onoga što čitaju, što može dovesti do nesporazuma, pogrešnih interpretacija i preteranih reakcija. Brzina kojom informacije cirkulišu stvara pritisak na pojedince da odmah reaguju, što ponekad vodi preuranjenim zaključcima i negativnom uticaju na javni diskurs i donošenje informisanih odluka.

Ova decentralizacija je dovela do fragmentacije medijskog prostora, gde pojedinci sve više borave u „eho komorama", okruženi istomišljenicima i informacijama koje potvrđuju njihove stavove, dok suprotstavljena mišljenja bivaju ignorisana ili iskrivljena. Eho komore i epistemički mehurovi predstavljaju fenomen u kojem pojedinci izbegavaju određene informacije ili izvore koji bi mogli osporiti njihova uverenja, čime se ograničava njihov pristup potpunijem znanju. Ovaj koncept se često koristi u kontekstu digitalnih medija, gde se korisnici selektivno izlažu informacijama koje su u skladu s njihovim vrednostima i stavovima. Jedan od ključnih radova koji razmatra eho komore i epistemičke mehurove je rad filozofa Ti Njujena (Thi Nguyen) pod nazivom „Eho komore i epistemički mehurovi" (Echo Chambers and Epistemic Bubbles) (2020). Njujen objašnjava kako epistemički mehurovi deluju kao filtrirajući mehanizmi, čineći da pojedinci veruju da su njihova shvatanja sveta potpuna, čak i kada su suštinski fragmentarna. Njujen navodi da „[...] internet tehnologije stvaraju hiperindividualizovane, tajne filtere. Tajnost je posebno preteća. Mnogi korisnici ne znaju za postojanje algoritamskog ličnog filtriranja" (Nguyen 2020: 5). On sugeriše da je ključna karakteristika epistemičkih mehurova nevoljnost ili nesposobnost pojedinca da bude izložen kontrastnim informacijama, što dalje doprinosi jačanju predrasuda i pogrešnih uverenja. Slično mišljenje iznosi i Majkl Linč (Michael Lynch) u radu „Internet zajednice" (The Internet of Us) (2016), gde ističe uticaj interneta na ljudsko znanje i epistemološke prakse. Linč ističe da „[...] s povećanom slobodom izražavanja i potrošnje dolazi rizik od povećane individualne izolacije" (Lynch 2016: 46). Ova pristrasnost u prikazivanju sadržaja može dodatno pojačati strukturnu polarizaciju jer su korisnici sve više izolovani u svojim grupama, a sadržaj koji im se prikazuje potvrđuje njihova već postojeća uverenja i stavove, umesto da ih izlaže raznolikim perspektivama, a takođe može

uticati i na nestrukturnu polarizaciju kroz personalizaciju sadržaja. Kada algoritmi analiziraju prethodno ponašanje korisnika i prilagođavaju sadržaj prema njihovim interesima, to može uticati na to kako se mišljenja i stavovi korisnika oblikuju kroz interakcije s drugima.

Na primer, ako se korisniku konstantno prikazuju slični ili polarizovani stavovi, može doći do promene u njegovim stavovima bez obzira na to da li je deo iste grupe ili ne. U osnovi, algoritamska pristrasnost može pospešiti oba oblika polarizacije – strukturnu i nestrukturnu – kroz način na koji korisnici komuniciraju i razmenjuju informacije, jer algoritmi koji favorizuju određene vrste sadržaja mogu pojačati razlike među grupama i uticati na to kako se pojedinci međusobno povezuju, čime se dodatno komplikuje dinamika mišljenja u digitalnom medijskom prostoru (Jacob & Banisch 2023). Dok su se u početku digitalni medijski sistemi oslanjali na ljudske moderatore za pregled i razvrstavanje sadržaja, sve veći broj informacija je doveo do pojave algoritama za moderaciju kao odgovor na potrebu čoveka da reguliše sve veći broj podataka na internetu. Kako je digitalni svet rastao, bilo je nemoguće održati ljudski faktor, te se digitalni svet suočio sa izazovima koji su bili van kapaciteta ljudske moći. Krajem prošlog i početkom novog veka pojavili su se prvi alati za automatsko filtriranje, kao što su spam filteri, koji su koristili jednostavne algoritme za identifikaciju i uklanjanje neželjenih poruka, a s kasnijim razvojem veštačke inteligencije i primenom mašinskog učenja algoritmi su postali sofisticiraniji, te su digitalne platforme poput Fejsbuka i drugih počele da koriste složene modele za analizu sadržaja, prepoznavanje obrasca i automatsku moderaciju. Ovi algoritmi su postali sposobni da prepoznaju ne samo neželjeni sadržaj već i govor mržnje, slike s neodgovarajućim sadržajem i dezinformacije (Gillespie 2018). Danas algoritmi za moderaciju igraju ključnu ulogu u održavanju bezbednosti i integriteta digitalnih zajednica.

Međutim, njihova upotreba dolazi sa izazovima, uključujući prepoznavanje konteksta, pristrasnost u podacima i rizik od cenzure. Kritičari ističu da algoritmi mogu nenamerno favorizovati određene stavove ili sadržaje, često nagrađujući materijale koji izazivaju snažne emotivne reakcije ili kontroverze, dok istovremeno marginalizuju umerenije ili kompleksnije perspektive. Jedan od najpoznatijih primera algoritamske pristrasnosti dogodio se u velikoj kompaniji Amazon,[2] koja je koristila sistem veštačke inteligencije prilikom zapošljavanja. Naime, u ovom slučaju je algoritam koristio sve podatke prilikom prijava za posao koje su podnete kompaniji tokom desetogodišnjeg

2 https://www.imd.org/research-knowledge/digital/articles/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human/, posećeno 10.09.2024.

perioda kako bi naučio kako da prepoznaje kvalifikacije najboljih kandidata. S obzirom na mali procenat žena koje rade u kompaniji, algoritam je brzo uočio mušku dominaciju, i smatrao je upravo to faktorom uspeha. Algoritam ovog sistema koristio je rezultate svojih vlastitih predviđanja kako bi poboljšao svoju preciznost, te se „zaglavio" u obrascu diskriminacije prema kandidatkinjama. Pošto su podaci korišćeni za obuku algoritma u nekom trenutku kreirani od strane ljudi, to znači da je algoritam takođe nasledio neželjene ljudske osobine, poput pristrasnosti i diskriminacije, koje su već godinama vodeći problemi u procesu zapošljavanja. Još jedan od primera koji će nam poslužiti je studija MIT Media Laba koja je pokazala da neki popularni sistemi veštačke inteligencije, uključujući one koje koriste velike tehnološke kompanije, imaju do 34 % greške u prepoznavanju lica žena tamnije puti, dok je ta greška kod muškaraca bele puti svega 0,8 %.[3]

Mašinsko učenje automatski otkriva obrasce u podacima i postalo je ključni alat u različitim tehnologijama, od pretraživača do sistema za sprečavanje prevara, uključujući primere kao što su digitalne kamere i automobili sa sistemima za sprečavanje nesreća. Zbog složenosti obrazaca koje mašine treba da prepoznaju, ljudski programeri ne mogu dati detaljna uputstva, pa mašinsko učenje omogućava programima da uče i prilagođavaju se na osnovu iskustva (Le 2021). Treniranje algoritama vrši se korišćenjem velikih skupova podataka kako bi se algoritmi naučili da prepoznaju obrasce i donose odluke na osnovu tih obrazaca. Ovaj proces je ključan u razvoju veštačke inteligencije i mašinskog učenja, tako da treniranje algoritama možemo opisati kroz faze prikupljanja podataka, zatim pripremu i na kraju testiranje.

Algoritmi na kojima se baziramo u ovom radu su algoritmi za moderaciju sadržaja koji se koriste za automatsko pregledanje, analizu i filtriranje sadržaja na digitalnim platformama, kao što su društvene mreže, forumi i komentari na veb sajtovima, a njihov cilj je da identifikuju i uklone sadržaje koji krše pravila platforme, uključujući nasilje, govor mržnje, dezinformacije, spamovanje ili nepristojan materijal. Pogrešna interpretacija algoritama za moderaciju odnosi se na situacije kada algoritmi netačno klasifikuju sadržaj zbog ograničenog razumevanja konteksta ili složenosti jezika. Na primer, algoritmi mogu označiti satiru, ironiju ili političku debatu kao uvredljiv sadržaj jer prepoznaju određene ključne reči bez razumevanja šireg značenja, a ova ograničenja često izazivaju frustraciju kod korisnika, jer algoritmi ne uspevaju da razlikuju štetne objave od onih koje su deo normalne društvene

3 https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212, posećeno 12. 09. 2024.

interakcije. Pored analize samog sadržaja, mogu pratiti ponašanje korisnika na platformi kako bi identifikovali sumnjive aktivnosti.

Na primer, korisnici koji postavljaju veliki broj komentara u kratkom vremenskom roku ili koriste ponavljajuće fraze mogu biti označeni kao botovi. Ako korisnik širi dezinformacije u mnogim grupama na društvenim mrežama u kratkom vremenskom periodu, algoritam može prepoznati to ponašanje kao sumnjivo i ograničiti mogućnost korisnika da objavljuje. Algoritmi često uklanjaju ili označavaju sadržaje koji ne krše pravila, ali ih prepoznaju kao nepoželjne zbog određene reči ili slike, te „[...] budućnost ekosistema društvenih medija možda već ukazuje na okruženja u kojima je interakcija između mašina norma, a ljudi će se kretati svetom koji većinom naseljavaju botovi. Verujemo da postoji potreba da botovi i ljudi mogu međusobno da se prepoznaju kako bi se izbegle bizarne ili čak opasne situacije zasnovane na lažnim pretpostavkama o ljudskim sagovornicima“ (Ferrara, Varol, Davis 2016: 8). To je naročito problem u kontekstu aktivističkih sadržaja ili složenijih diskusija koje koriste jezik ili slike koje algoritmi pogrešno interpretiraju. Pristrasnost algoritama odnosi se na tendenciju algoritama da favorizuju određene rezultate ili odluke zbog inkorporiranih parametara u podacima, u samom dizajnu ili načinu korišćenja. Kako algoritmi određuju koji sadržaj postaje vidljiviji korisnicima, postoji rizik da favorizuju određene ideologije ili interese, potiskujući različita mišljenja. Ova dinamika može doprineti jačanju stereotipa i predrasuda, otežavajući dijalog i razumevanje među različitim grupama. Na primer, u istraživanju objavljenom 2020. godine (Hofmann et al. 2024) pokazano je da su postovi Afro-Amerikanaca na društvenim mrežama češće uklanjani ili označavani kao neprikladni, jer su algoritmi uočili određene obrasce jezika ili tema koje su pogrešno prepoznate kao prekršaji. Ključni nalaz istraživanja je da jezički modeli održavaju oblik prikrivenih rasnih predrasuda prema Afro-Amerikancima aktiviranih dijalektom, dok su ostale anketirane zajednice otvoreno pokazivale pozitivne stavove prema ovoj zajednici, njihovo nesvesno ponašanje podržavalo je rasne nejednakosti. Negativni stereotipi opstaju čak i kada su površno odbacivani, a jezički modeli reprodukuju arhaične stereotipe, pokazujući ideologiju bez boje (Hofmann et al. 2024). „[…] Istorijski podaci banke o kreditiranju mogu pokazati da je ona rutinski i nepravedno davala veće kamatne stope stanovnicima u većinski crnačkom delu grada prema poštanskom kodu. Algoritam za bankarstvo obučen na tim pristrasnim podacima mogao bi prepoznati taj obrazac diskriminacije i naučiti da naplaćuje stanovnicima u tom delu grada, a prema poštanskom kodu, više za kredite, čak i ako ne zna rasu podnosioca zahteva.“ (Le 2021: 4). Algoritamska pristrasnost nastaje kada algoritmi do-

nose nepravedne odluke koje privileguju određene grupe, i svest o ovome je veoma značajna zato što se koriste za donošenje odluka u oblastima poput zapošljavanja, obrazovanja, stanovanja i kreditiranja, što može uticati na ekonomske prilike i doprineti produbljivanju nejednakosti (Le 2024: 5).

Programeri koji kreiraju algoritme moraju da balansiraju između maksimizacije profita i nediskriminacije kroz nekoliko segmenata, kao što su definisanje ciljeva algoritma, korišćenje transparentnih podataka i temeljno testiranje. Oni su dizajnirani da optimizuju određene ciljeve, kao što su profit, tačnost ili korisnička pozitivna potvrda, te uvođenjem nediskriminacije kao jednog od ciljeva u funkciji algoritma, programeri mogu osigurati da ne donose odluke koje favorizuju određene grupe na nepravedan način. Ako podaci koji se koriste sadrže pristrasnost (npr. istorijska diskriminacija), algoritam će verovatno inkorporirati tu pristrasnost. Istorijska diskriminacija inkorporirana u algoritme označava način na koji prethodne predrasude i nejednakosti, koje su se dogodile u društvu, utiču na dizajn i funkcionalnost algoritama. Ovo se dešava kada se oni obučavaju na osnovu istorijskih podataka koji odražavaju sistemsku diskriminaciju, kao što su rasna nejednakost, određeni pol, etnička pripadnost ili socioekonomski status. Kao rezultat, algoritmi mogu nesvesno perpetuirati te pristrasnosti, favorizovati određene grupe i potiskivati druge, čime se dodatno produbljuju postojeće nejednakosti. Ova dinamika može dovesti do nepravednih odluka u različitim oblastima, uključujući zapošljavanje, pravdu, zdravstvo i kreditiranje, jer algoritmi ne prepoznaju ili ne ispravljaju istorijske nepravednosti, već ih reprodukuju u savremenim kontekstima.

U mnogim zemljama postoje zakoni i regulative koji zabranjuju diskriminaciju, te programeri moraju osigurati da algoritmi poštuju ove zakonske okvire dok se i dalje teži ispunjenju ostalih funkcija algoritama. Takođe, mnoge kompanije usvajaju etičke kodekse kako bi se obezbedila pravednija upotreba tehnologija. Na ovaj način, programeri moraju pažljivo balansirati između komercijalnih ciljeva i etičkih principa, te se njima preporučuje da razumeju različite vrste pristrasnosti, kako bi prepoznali potencijalne izvore pristrasnosti u podacima, da pažljivo analiziraju izvor i demografske karakteristike podataka koje koriste, kao i da osiguraju raznolikost unutar timova kako bi uključili različite perspektive u sam razvoj sistema. Transparentnost u procesima odlučivanja i dobro dokumentovanje metodologije, zatim redovno testiranje i evaluacija na osnovu različitih demografskih grupa može pomoći u identifikaciji i ispravljanju pristrasnosti, čime se doprinosi pravednijim i inkluzivnijim rešenjima.

Još jedan od primera neetičnosti primenjenih algoritama jeste slučaj engleske Kancelarije za kvalifikacije i regulaciju ispita Ofkval (Ofqual), koja je koristila sistem za izračunavanje ocena učenika. Da bi izračunao ocene, algoritam se oslanjao na predviđanja nastavnika o završnim ocenama učenika, njihov akademski uspeh i, što je najvažnije, istorijske podatke o renomeu škole iz prethodnih godina. Algoritam je snizio 40 % ocena koje su dali nastavnici u procesu izračunavanja konačnih rezultata, a analiza sistema je pokazala da je algoritam verovatnije snižavao ocene učenicima iz nižih socioekonomskih slojeva i onima koji nisu pohađali privatne škole. Nakon velikog javnog negodovanja, Ofkval je povukao algoritamske ocene i učenici su dobili ocene koje su im dali njihovi nastavnici. Ovo je još jedan primer nepodudaranja između ishoda koji algoritam treba da predvidi i onoga što zapravo predviđa, zato što nije određivao stvarno postignuće učenika tokom godine, već je predviđao koliko bi učenici u određenoj školi „trebalo" da postignu. Fokus algoritma na renome škole kao prediktor značio je da su učenici s visokim postignućima u lošije rangiranim školama imali veću verovatnoću da im ocene budu snižene. Pored toga, algoritam je davao veći značaj ocenama nastavnika u školama s manjim brojem učenika, na taj način dajući učenicima iz privatnih škola nepravednu prednost (Le 2021: 17).

U radu „Algoritamska pristrasnost društvenih mreža" (Algorithmic Bias of Social Media) Daman Prit Singa (Daman Preet Singh) autor istražuje problem pristrasnosti u algoritmima društvenih mreža i njihove posledice na korisničko iskustvo i društvo u celini. Sing objašnjava šta podrazumeva algoritamska pristrasnost, naglašavajući da se ona javlja kada algoritmi favorizuju određene grupe ili vrste sadržaja na osnovu istorijskih podataka ili namerno postavljenih kriterijuma. Ova pristrasnost može biti rezultat podataka koji su prikupljeni u prošlosti, i koji često odražavaju društvene predrasude. Takođe analizira kako algoritmi oblikuju informacije koje korisnici vide na platformama poput Fejsbuka, Tvitera i Instagrama. Pristrasni algoritmi mogu stvoriti „eho komore", gde korisnici vide sadržaj koji je u skladu s njihovim prethodnim interakcijama, a ne raznovrsne perspektive. Ovo može dovesti do smanjenja kritičkog mišljenja i jačanja polarizacije, te on ističe da pristrasnost u algoritmima može doprineti jačanju stereotipa, dezinformaciji i marginalizaciji određenih grupa.

Na primer, algoritmi mogu favorizovati sadržaj koji odražava negativne stereotipe o određenim etničkim ili socijalnim grupama, čime se dodatno produbljuju predrasude u društvu. U radu se bavi i etičkim dilemama koje se pojavljuju u vezi s razvojem algoritama, i poziva na veću odgovornost pro-

gramera i kompanija koje razvijaju ove sisteme da razmotre društvene posledice svojih odluka (Singh 2023).

U stvarnosti, algoritmi koji favorizuju kontroverznu i polarizovanu sadržinu često generišu veći angažman korisnika, što dovodi do povećane interakcije, a samim tim i do većih prihoda od oglašavanja. Ova situacija stvara etičku dilemu, jer kompanije moraju balansirati između svoje odgovornosti prema društvu i svoje potrage za profitom, te dok se neki pozivaju na veći nivo odgovornosti i transparentnosti u načinu na koji algoritmi funkcionišu, kompanije često minimalno reaguju na pritiske javnosti, bez stvarne promene u osnovnim praksama koje bi mogle umanjiti polarizaciju. Osim toga, nedostatak pravne regulative i standarda u industriji dodatno komplikuje situaciju, jer različite platforme imaju različite pristupe u vezi s pitanjima sigurnosti i etike.

Bez jasnih smernica i okvira za delovanje, kompanije mogu nastaviti s praksama koje donose kratkoročne profite, dok dugoročno štete društvenom tkivu. Zbog toga su dosadašnji napori ka samoregulaciji često propadali, a etičke smernice za veštačku inteligenciju ponekad se nazivaju „papirnim tigrovima".[4] Možda će biti potrebno da zakoni ili regulatorna tela nametnu ove promene, a budući da su u suprotnosti s poslovnim modelima koji teže monopolizaciji pažnje, biće ih teško sprovesti u delo. Da bi se smanjila pogrešna predviđanja algoritama za moderaciju, koja za rezultat imaju polarizaciju ili diskriminaciju, ključno je uvesti balans između automatizacije i ljudske kontrole. Transparentnost u radu algoritama, kao i njihova objašnjivost, pomoći će korisnicima da razumeju kako se donose odluke o moderaciji. Takođe, algoritme treba prilagoditi za prepoznavanje kulturnih i jezičkih varijacija, a povratne informacije korisnika koristiti za stalno unapređenje. Implementacija ovih preporuka za smanjenje pogrešnih predviđanja algoritama za moderaciju jeste realna, ali dolazi sa izazovima. Tehnički gledano, većina tih rešenja je moguća – algoritmi se mogu trenirati na raznovrsnim podacima, može se uključiti ljudska provera i poboljšati transparentnost. Međutim, u praksi izazovi uključuju resurse, vreme i finansije potrebne za stalnu nadogradnju sistema, kao i balansiranje između efikasnosti i pravednosti, kao i prepoznavanje takvih pretnji u vidu loših predviđanja i njihovog uticaja na društvo.

4 Jye Beardow je student na Australijskom nacionalnom univerzitetu. Verzija ovog rada je napravljena za LAWS4283, za koju je Jye dobio *Proximity* nagradu za informatičko pravo. Pogledati na www.24469-scroll-click-like-share-repeat-the-algorithmic-polarisation-phenomenon%20(1).pdf.

Poštovanje etičkih standarda u treniranju algoritama zahteva pristup koji obezbeđuje da oni funkcionišu bez pristrasnosti prema određenim grupama, što se postiže korišćenjem raznovrsnih i reprezentativnih podataka koji reflektuju različite demografske i kulturne karakteristike. U radu „Pravednost i apstrakcija socio-tehničkih sistema" (Fairness and Abstraction in Sociotechnical Systems) (Selbst et al. 2019) autori istražuju izazove treniranja algoritama na način da konačni ishod bude etičan, naglašavajući da trenutni pristupi često zanemaruju širi društveni kontekst, što može dovesti do posledica polarizacije i diskriminacije. Identifikuju pet „zamki" u ovoj oblasti i to: zamka okvirnog postavljanja (Framing Trap), koja ukazuje na važnost poznavanja šireg društvenog konteksta, zatim zamka prenosivosti (Portability Trap), gde se algoritmi neadekvatno primenjuju u novim kontekstima, zamka formalizma (Formalism Trap), koja redukuje pravednost na stroge definicije, zamka efekta talasa (Ripple Effect Trap), koja ukazuje na nepredviđene posledice tehnologija, i zamka rešenja (Solution Trap), koja se fokusira na tehnička rešenja bez razumevanja društvenih potreba. Autori naglašavaju potrebu da dizajneri algoritama razviju svest o društvenim i političkim kontekstima, kao i da sarađuju s relevantnim stručnjacima iz drugih oblasti, čime se stvara osnova za postizanje pravednijih i održivijih rešenja.

Bitno je identifikovati i ukloniti predrasude u podacima i modelima, uz implementaciju mehanizama za redovno praćenje i prilagođavanje, kako bi se osigurala transparentnost i odgovornost. Uključivanje različitih perspektiva i etičkih standarda u proces razvoja takođe doprinosi stvaranju pravednijih algoritama koji mogu smanjiti nejednakosti i povećati poverenje korisnika. Osim toga, potreban je značajan trud za osiguranje transparentnosti i etičkog nadzora, što zahteva resurse i obuku zaposlenih. Dok veće kompanije često imaju kapacitete za implementaciju ovih praksi, manje organizacije se mogu suočiti sa ograničenjima u pristupu tehnologiji i ekspertizi. U tom smislu, iako je pravednost u treniranju algoritama moguća, njena primena zahteva kontinuirani rad i posvećenost, kao i saradnju između različitih disciplina i stručnjaka. Preporuke za sprečavanje nepredviđenih rezultata algoritamske moderacije uključuju sistematsko razumevanje različitih kulturnih aspekata i metajezika društvenih zajednica u kojima se algoritmi primenjuju, kako bi se obezbedile sve relevantne perspektive. Pravne regulative bi trebalo da obuhvate mehanizme za nezavisnu reviziju i praćenje algoritamskih odluka, kao i uspostavljanje odgovornosti platformi za potencijalne negativne posledice na društveni dijalog i pluralizam mišljenja. Time bi se obezbedio balans između efikasnosti moderacije i zaštite osnovnih prava korisnika, kao što su sloboda izražavanja i pristup različitim stavovima.

Pre šire primene algoritama neophodno je sprovesti pilot-testiranja u kontrolisanim uslovima kako bi se procenile njihove posledice, što se nameće kao obaveza velikim tehnološkim kompanijama koje te sisteme razvijaju. Takođe, važno je razvijati algoritme koji se mogu prilagoditi kulturnim promenama, kao i uključiti etičke smernice u njihov dizajn, čime se smanjuje rizik od pristrasnosti i diskriminacije, što zajedno doprinosi pravednijim i inkluzivnijim ishodima. Države su u obavezi da razvijaju regulative koje zahtevaju transparentnost i odgovornost tehnoloških kompanija, dok građanski aktivizam može stvoriti pritisak za poboljšanje ovakvih praksi, jer „[...] kako algoritamska moderacija postaje sve više integrisana u svakodnevno onlajn iskustvo korisnika, te zaštitnici ljudskih prava i istraživači moraju nastaviti da preispituju i diskurs i stvarnost korišćenja automatizovanog donošenja odluka u moderaciji, ne dozvoljavajući firmama da se skrivaju iza vela složene crne kutije dok pokušavaju da se udalje od važnih diskusija o samoj politici sadržaja“ (Gorwa, Binns, Katzenbach 2020).

Saradnja sa akademskom zajednicom pomaže u razvijanju pravednijih rešenja i praćenja etičnosti, dok formiranje etičkih okvira osigurava da se lokalne kulturne potrebe uzimaju u obzir. Takođe, podsticanje otvorenih i decentralizovanih platformi može smanjiti zavisnost od velikih tehnoloških kompanija, čime se ukida monopol nad razvijanjem ovih sistema, što bi uključivalo i značajne finansijske resurse samih država. Kombinovanjem ovih pristupa društvene zajednice mogu aktivno raditi na umanjivanju negativnih posledica algoritamskih rešenja i osiguravanju pravednijih ishoda u njihovoj primeni.

Algoritmi su alati koje razvijaju i implementiraju ljudi, a njihovo funkcionisanje odražava ljudske odluke, vrednosti i predrasude. Iako algoritmi mogu automatski analizirati velike količine podataka i pružiti preporuke ili sadržaj, konačna odgovornost za to kako se ti alati koriste i koji se sadržaji promovišu leži na ljudima koji ih dizajniraju i upravljaju njima. Stoga je za stvaranje zdravijeg i pravednijeg digitalnog okruženja i smanjenja uticaja digitalnih medijskih sistema na društvenu polarizaciju neophodno da se javnost aktivno uključi u oblikovanje pravila, smernica i etičkih standarda koji će voditi korišćenje tehnologije, te „[...]da bismo spasili naše digitalno okruženje od samog sebe, na kraju ćemo morati da formiramo novu grupu digitalnih ekologa – građana ovog novog prostora koji svi zajedno gradimo, koji će se udružiti kako bi zaštitili ono što je najbolje u njemu“ (Pariser 2011: 74).

To podrazumeva promenu u načinu razmišljanja o algoritmima, prepoznavanje njihove uloge u oblikovanju društvenog diskursa i aktivno traženje reše-

nja koja će osigurati da tehnologija služi ljudima, a ne obrnuto. Na taj način, možemo stvoriti informativno okruženje koje podstiče raznolikost mišljenja, kritičko razmišljanje i inkluzivnost, umesto da produbljuje podele. Kada postanemo svesni načina na koji algoritmi oblikuju našu percepciju i odlučivanje, stičemo moć da zahtevamo odgovornost od kompanija koje upravljaju ovim tehnologijama. Ova svest može poslužiti kao osnova za vođenje konstruktivnog dijaloga o etici, transparentnosti i pravednosti u korišćenju tehnologije, čime se otvaraju vrata za sistemske promene koje će doprineti unapređenju društvene kohezije. Na taj način naše znanje i razumevanje mogu postati ključ za izgradnju boljeg i harmoničnijeg društva, gde se različita mišljenja poštuju, a dijalog se podstiče.

Literatura

- Barocas, S., & Selvaraj, M. (2016). *Big data's disparate impact. Proceedings of the 2016 Conference on Fairness, Accountability, and Transparency (FAT*)*, 1–20. https://doi.org/10.1145/3287560.3287598.
- Bennett, W. L., & Iyengar, S. (2008). "A new era of minimal effects? Revisiting the impact of news media on political campaigns." *Journal of Communication*, 58(4), 707–731. DOI: https://doi.org/10.1111/j.1460-2466.2008.00457.x.
- Chomsky, N. (2002). *Media Control: The Spectacular Achievements of Propaganda*. Seven Stories Press.
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
- Ferrara, E., Varol, O., Davis, C. A., Menczer, F., & Flammini, A. (2016). "The rise of social bots." *Communications of the ACM*, 59(7), 96–104.
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). "Algorithmic content moderation: Technical and political challenges in the automation of platform governance." *Big Data & Society*, 7(1).
- Herman, E. S., and Chomsky, N. (1988). *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon Books.
- Hofmann, V., Kalluri, P.R., Jurafsky, D. et al. (2024) "AI generates covertly racist decisions about people based on their dialect." *Nature* 633, 147–154. DOI: https://doi.org/10.1038/s41586-024-07856-5.
- Jacob, D. & Banisch, S. (2023). "Polarization in Social Media: A Virtual Worlds-Based Approach." *Journal of Artificial Societies and Social Simulation*, 26. DOI: 10.18564/jasss.5170.

- Jenkins, H. (2006). *Convergence Culture: Where Old and New Media Collide*. New York University Press.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2016 Conference on Fairness, Accountability, and Transparency*, 1–16.
- Le, V. (2021). How *automated decision-making becomes automated discrimination*. Greenlining Institute. https://greenlining.org/wp-content/uploads/2021/04/Greenlining-Institute-Algorithmic-Bias-Explained-Report-Feb-2021.pdf.
- Lynch, M. P. (2016). *The internet of us: Knowing more and understanding less in the age of big data*. Liveright Publishing.
- Lipton, Z. C. (2016). *The mythos of model interpretability. Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 96–100. https://arxiv.org/abs/1606.03490.
- Maggino, F. & Fattore, M. (2019). *Social Polarization*. DOI: 10.1002/9781118445112.stat08138.
- Nordbrandt, M. (2021). "Affective polarization in the digital age: Testing the direction of the relationship between social media and users' feelings for out-group parties." *New Media & Society*, 25. DOI: 10.1177/14614448211044393.
- Nguyen, C. T. (2020). "Echo chambers and epistemic bubbles." *Episteme*, 17(2), 141–161. DOI: https://doi.org/10.1017/epi.2018.32.
- Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press.
- Radojković, M. (2017). „Digitalni mediji u Srbiji – Koristi i opasnosti". *Politeia*, 7(13).
- Selbst, A., Boyd, D., Friedler, S., Venkatasubramanian, S. & Vertesi, J. (2019). "Fairness and Abstraction in Sociotechnical Systems." 59-68. DOI: 10.1145/3287560.3287598.
- Singh, D. (2023). "Algorithmic Bias of Social Media." *The Motley Undergraduate Journal*. 1. DOI: 10.55016/ojs/muj.v1i2.77457.
- Sweeney, L. (2013). "Discrimination in online ad delivery." *Communications of the ACM*, 56(5), 44-54. DOI: https://doi.org/10.1145/2460276.2460278.
- Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D. & Nyhan, B. (2018). "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3144139.
- Vosoughi, S., Roy, D., & Aral, S. (2018). "The spread of true and false news online." *Science*, 359(6380), 1146–1151.

VEBOGRAFIJA

- https://www.imd.org/research-knowledge/digital/articles/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human/, posećeno 10.09.2024.
- https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212, posećeno 12.09.2024.
- Jye B. Pogledati na www.24469-scroll-click-like-share-repeat-the-algorithmic-polarisation-phenomenon%20(1).pdf.

Katarina Šmakić
Faculty of Diplomacy and Security, Belgrade

THE IMPACT OF DIGITAL MEDIA SYSTEMS ON SOCIAL POLARIZATION

Abstract

Moderation algorithms have become an essential part of contemporary digital media systems, enabling automatic filtering and removal of inappropriate content, thus eliminating the need for human intervention. This paper explores the implications of such an approach, with a particular focus on its impact on polarization within society. While algorithms provide efficiency in suppressing violent and inappropriate content, their operation is based on predefined rules that may unintentionally filter out certain viewpoints and/or information. This process can enhance the creation of closed information loops and narrow interest communities, as algorithms favour content that aligns with users' previous interests, reducing their exposure to opposing views and contributing further to polarization. The conclusion of this paper emphasizes the need to develop sophisticated mechanisms for recognizing algorithmic moderation that contributes to polarization. A key recommendation is to establish a legal framework that mandates digital platforms to ensure transparency in their algorithms, including the publication of criteria by which content is filtered and moderated, as well as the data on which they are trained. Legal regulations should encompass mechanisms for independent review and monitoring of algorithmic decisions, as well as establish accountability for platforms regarding potential negative impacts on social dialogue and the plurality of opinions. This would ensure a balance between the efficiency of moderation and the protection of fundamental user rights, such as freedom of expression and access to diverse viewpoints.

Keywords

digital media systems, algorithmic moderation, transparency, dialogue, freedom of expression.

Primljeno: 14. 09. 2024.
Prihvaćeno: 6. 10. 2024.