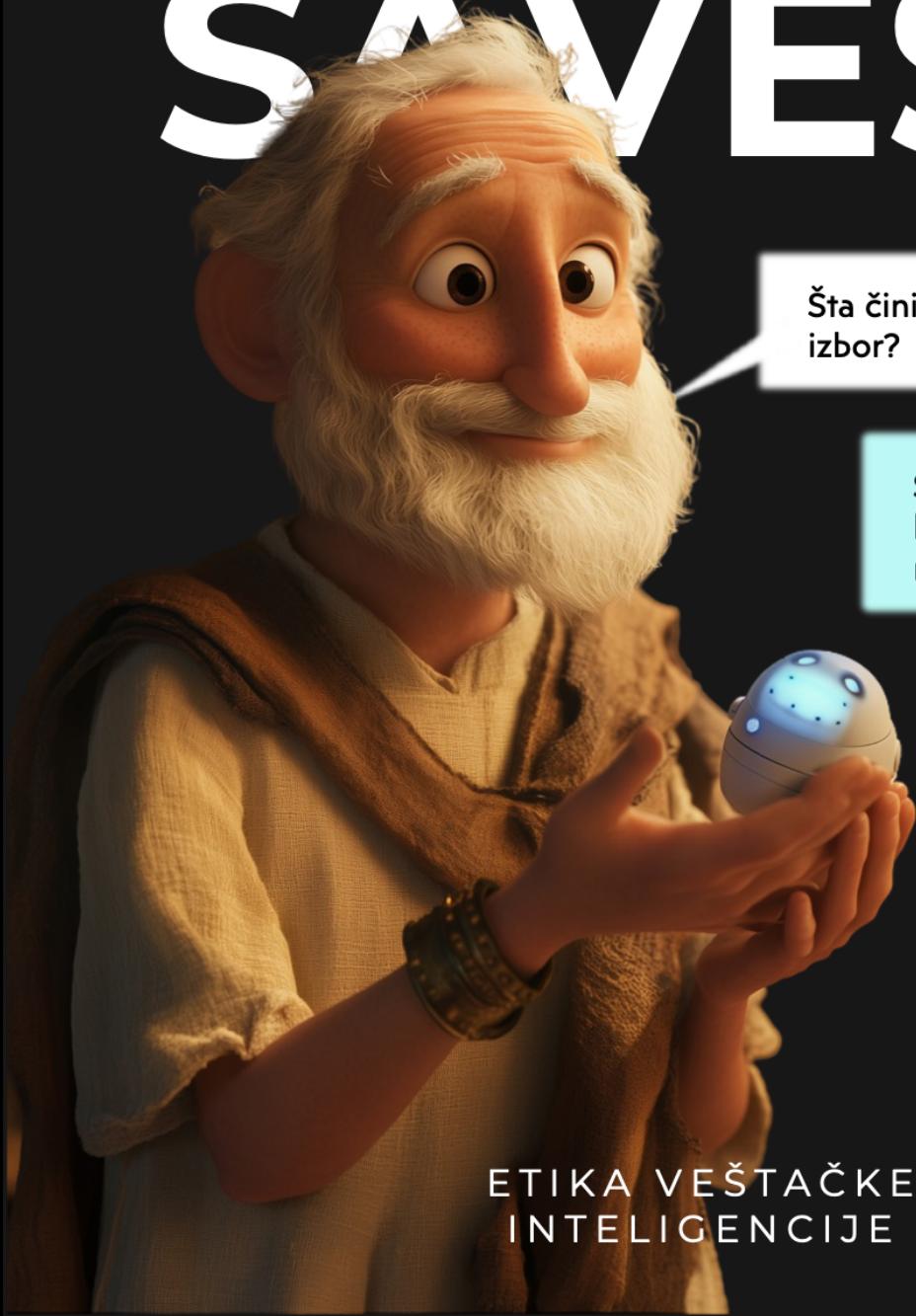


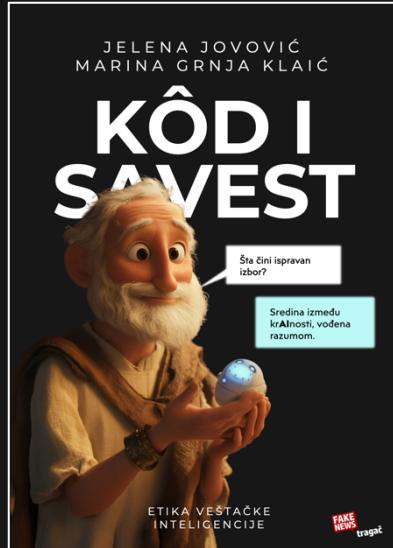
JELENA JOVOVIĆ
MARINA GRNJA KLAIĆ

KÔDI SAVEST



ETIKA VEŠTAČKE
INTELIGENCIJE

FAKE
NEWS
tragač



KÔD I SAVEST – Etika veštačke inteligencije

Novosadska novinarska škola
FakeNews Tragač
Kosovska 1,
2100 Novi Sad
Telefon: 021/ 424246
Mail: office@novinarska-skola.org.rs

Za izdavača
Milan Nedeljković

Autori
Jelena Jovović
Marina Grnja Klać

Dizajn
Teodora Koledin

Novi Sad,
Februar 2025.

JELENA JOVOVIĆ
MARINA GRNJA KLAIĆ

KÔD I **SAVEST**

ETIKA VEŠTAČKE
INTELIGENCIJE

U poslednjih nekoliko godina svedoci smo ekspanzivnog razvoja i upotrebe sistema veštačke inteligencije. Na primer, najpoznatiji četbot Čet Dži-Pi-Ti (*ChatGPT*) koji je lansiran u novembru 2022. godine je već početkom 2025. godine imao 300 miliona aktivnih korisnika širom sveta (1). U isto vreme svedočimo da trgovci uvode sisteme nadzora zasnovane na veštačkoj inteligenciji kako bi se zaštitali od krađa kupaca (2), u medicini se već primenjuje niz alata koji asistiraju lekarima u uspostavljanju dijagnoza različitih bolesti (3), veštačka inteligencija se koristi za analizu podataka u ratnim konfliktima (4, 5), kontinuirano se razvijaju novi algoritmi za društvene mreže koji omogućavaju efikasniju isporuku sadržaja i oglasa (6), a do sredine 2024. godine, besplatno ili uz nadoknadu manju od 5 dolara mesečno, i laicima i ekspertima je na raspolaganju bio značajan broj alata veštačke inteligencije – 2,298 alata za zamenu lica, sinhronizaciju usana za audio zapisom, animaciju lica i kreiranje AI avatara, 10,206 alata za generisanje slika, 1,018 alata za generisanje ili kloniranje glasa (7).

Ove brze promene, osim koristi za društvo i pojedince nose i značajne rizike, te se otvaraju ozbiljna i brojna etička pitanja koja se odnose na razvoj i primenu veštačke inteligencije. Pitanja koja se nameću odnose se na širok spektar oblasti, uključujući donošenje odluka, rad i zapošljavanje, društvene interakcije, zdravstvenu zaštitu, obrazovanje, medije, pristup informacijama, zaštitu privatnih podataka, zaštitu potrošača, životnu sredinu, demokratiju, vladavinu prava, bezbednost, kao i ljudska prava i osnovne slobode, poput slobode izražavanja, privatnosti i nediskriminacije. Imajući u vidu da pojedini sistemi veštačke inteligencije preuzimaju zadatke koje su ranije izvršavali ljudi i ulaze interakciju sa ekosistemima i okruženjem stvara se i novi društveni kontekst u kojem odrastaju nove generacije. U dugoročnoj perspektivi, sistemi veštačke inteligencije mogu uticati na način na koji ljudi shvataju sopstveno iskustvo i delovanje, kako doživljavaju i razumeju sebe, što otvara dodatna pitanja o društvenim, kulturnim i ekološkim interakcijama, autonomiji, sposobnosti delovanja, kao i o ljudskoj vrednosti i dostoanstvu (8).

Nameru nam je da se u ovom radu upravo fokusiramo na etiku veštačke inteligencije, posebno one generativne, sa čijom (zlo)upotrebom se fektčekeri u svom radu najčešće susreću.

Šta je veštačka inteligencija?

U stručnoj literaturi nailazimo na brojne definicije veštačke inteligencije a ono što ih povezuje jeste naglasak na sposobnosti mašina da obavljaju određene uloge i zadatke koje trenutno izvršavaju ljudi, kako na radnom mestu, tako i u društvu uopšte

(9). Ekspertska grupa za veštačku inteligenciju koja je formirana u okviru Evropske komisije 2019. godine pod pojmom veštačka inteligencija podrazumeva „sisteme koji ispoljavaju intelligentno ponašanje analizirajući svoje okruženje i preduzimajući akcije – s određenim stepenom autonomije – kako bi ostvarili određene ciljeve“, pri čemu ovi sistemi mogu biti isključivo softverski, delovati u virtuelnom svetu (npr. glasovni asistenti, softveri za analizu slika, pretraživači, sistemi za prepoznavanje govora i lica) ili mogu biti ugrađeni u hardverske uređaje – kao što su napredni roboti, autonomna vozila, dronovi ili aplikacije Interneta stvari (10).

Prepoznujući stalni tehnološki razvoj u ovoj oblasti i potrebu da i definicije usklađuju sa tim razvojem, Organizacija Ujedinjenih nacija za obrazovanje, nauku i kulturu (UNESCO) u svojim Preporukama o etici veštačke inteligencije, definiše veštačku inteligenciju uzimajući u obzir one aspekte koji su relevantni za etička pitanja. Stoga se sa stanovišta ovih Preporuka sistemi veštačke inteligencije definišu kao sistemi koji imaju sposobnost obrade podataka i informacija na način koji podseća na ljudsko intelligentno ponašanje, obuhvatajući aspekte poput rezonovanja, učenja, percepcije, predikcije, planiranja i upravljanja. Tri ključna elementa su u središtu ovog pristupa:

1. Sistemi veštačke inteligencije su tehnologije za obradu informacija koje integrišu modele i algoritme, omogućavajući sposobnost učenja i obavljanja kognitivnih zadataka, što dovodi do rezultata kao što su predviđanje i donošenje odluka u materijalnim i virtuelnim okruženjima. Ovi sistemi su dizajnirani da funkcionišu sa različitim stepenima autonomije, koristeći modelovanje i reprezentaciju znanja, kao i analizu podataka i izračunavanje korelacija.
2. Sistemi veštačke inteligencije mogu uključivati metode mašinskog učenja, uključujući duboko učenje i podržano učenje.
3. Sistemi veštačke inteligencije mogu uključivati metode mašinskog rezonovanja, uključujući planiranje, rasporedivanje, reprezentaciju i obradu znanja, pretragu i optimizaciju (11).

Generativna veštačka inteligencija, na koju se fokusiramo u ovom radu, je krovni termin koji se odnosi na sisteme veštačke inteligencije koji su sposobni da kreiraju tekst, slike ili druge oblike medija korišćenjem generativnih modela. Ovi modeli uče obrasce i strukture u okvirima podataka koji su korišćeni za njihovu obuku, a zatim generišu nove podatke sa sličnim karakteristikama. Razlikujemo različite tipove generativne veštačke inteligencije (kao što su generativne adversarijske mreže – GAN, modele zasnovane na transformatorima – TRM, varijacione autoenkodere – VAI, i dr.) od kojih je svaki kreiran tako da izvršava različite zadatke, odnosno da generiše različite tipove medija (12).

Zahvaljujući dostupnosti i jednostavnosti upotrebe, ovi sistemi se sve više i upotrebljavaju i zloupotrebljavaju. U petogodišnjem periodu od 2019. do 2023. godine, u digitalnom okruženju je samo broj videa koji su generisani veštačkom inteligencijom porastao za 550% i dospio je broj od 95,820 ([13](#)). U 2024. godini, polovina kompanija širom sveta prijavila je incidente povezane sa dipfejk prevarama, među kojima se po smelosti izdvaja ona u kojoj su prevaranti upriličili video-konferencijski poziv sa radnikom jedne multinacionalne kompanije u Hong Kongu na kojem su mu dipfejk kolege, uključujući i „finansijskog direktora“, naložile da im sa računa kompanije isplati 25 miliona dolara ([14](#)).

Etika i regulatorni okviri

Prvi globalni etički okvir veštačke inteligencije sadržan je u već pomenutom dokumentu *Preporuke o etici veštačke inteligencije* koji su u novembru 2021. godine usvojile 193 zemlje članice UNESKO-a. U ovim preporukama etika se posmatra kao „dinamična osnova za normativnu evaluaciju i usmeravanje tehnologija veštačke inteligencije, oslanjajući se na ljudsko dostojanstvo, dobrobit i prevenciju štete kao kompas, ukorenjen u etici nauke i tehnologije“ (str. 9, [8](#)).

Cilj ovih *Preporuka* je uspostavljanje univerzalnog okvira vrednosti, principa i smernica koji će podržati države u razvoju zakonodavstva, politika i drugih regulatornih instrumenata za tehnologije veštačke inteligencije u skladu sa međunarodnim pravom. Preporuke nastoje da integrišu etičke aspekte u sve faze životnog ciklusa sistema veštačke inteligencije, uz zaštitu, promociju i poštovanje ljudskih prava, dostojanstva i jednakosti, uključujući rodnu ravnopravnost.

Pored toga, *Preporuke* teže da zaštite interes sadašnjih i budućih generacija, očuvanju životne sredine, biodiverziteta i ekosistema, kao i očuvanju kulturne raznolikosti u razvoju i primeni tehnologija veštačke inteligencije. Naglašava se važnost multidisciplinarnog i inkluzivnog dijaloga o etici veštačke inteligencije i promoviše pravičan pristup u razvoju i deljenju znanja u ovoj oblasti, sa posebnim fokusom na potrebe i doprinose zemalja sa nižim prihodima, uključujući najmanje razvijene zemlje.

Oslanjajući se na obaveze koje proizilaze iz usvajanja ovog dokumenata, kao i niza drugih međunarodnih sporazuma i dokumenata, u martu 2023. godine Vlada Republike Srbije usvojila je *Etičke smernice za razvoj, primenu i upotrebu pouzdane i odgovorne veštačke inteligencije*. Osnovna načela koja se izdvajaju u *Etičkim smernicama* su:

- *Objašnjivost i proverljivost* koje podrazumevaju transparentnost u procesima razvoja, testiranja puštanja u rad, nadzora nad radom i gašenjem sistema veštačke inteligencije te njihovu proverljivost u svim ovim procesima, uključujući njihove kratkoročne i dugotrajne uticaje na čoveka.

- *Dostojanstvo* koje se određuje kao polazno načelo koje u fokusu ima očuvanje ljudskog integriteta i naglašava da svaki pojedinac ima pravo na zaštitu sopstvenog *dostojanstva*, a kršenje ili nepoštovanje ovog prava se sankcioniše zakonom.

- Načelo *zabrane činjenja štete* podrazumeva da sistem veštačke inteligencije mora biti u skladu sa bezbednosnim standardima, odnosno treba da sadrži odgovarajuće mehanizme koji će sprečiti bilo kakvu štetu po ljude i njihovu imovinu. Ukoliko do štete dođe, ona mora biti otklonjena u najkraćem mogućem roku, a oštećeno lice treba da bude obeštećeno u skladu sa zakonom.

- Načelo *pravičnosti* podrazumeva zaštitu od diskriminacije, posebno diskriminacije ranjivih društvenih kategorija. Ono uključuje stvarnu dimenziju koja podrazumeva zaštitu od neopravdane pristrasnosti, diskriminacije i stigmatizacije, kao i proceduralnu - koja uključuje mogućnost osporavanja i pružanje efikasne pravne zaštite protiv odluka koje su rezultat rada sistema veštačke inteligencije, kao i od osoba odgovornih za rad sistema ([15](#)).

Iako se prilikom određenja načela dostojanstva i zabrane činjenja štete *Etičke smernice za razvoj, primenu i upotrebu pouzdane i odgovorne veštačke inteligencije* pozivaju na neprikošnovenost ljudskog dostojanstva garantovanu Ustavom, *Zakon o javnom informisanju i medijima* koji propisuje da je dostojanstvo ličnosti lica na koji se odnosi informacija pravno zaštićeno, štiti dostojanstvo ličnosti, zatim *Zakon o obligacionim odnosima* koji uređuje pojam štete, *Krivični zakonik* koji propisuje sankcije za vredanje ljudskog dostojanstva i predviđa krivična dela protiv života i tela, imovine ljudi i sloboda i prava čoveka u domenu zabrane činjenja štete, one ipak nemaju ambiciju da regulišu ovu oblast, već samo prepoznaju potrebu za njenim zakonskim uređenjem.

Prvi sveobuhvatni pravni okvir o veštačkoj inteligenciji, u obliku *Zakona o veštačkoj inteligenciji Evropske unije*, stupio je na snagu u februaru 2025. godine. Kako stoji u prvom članu, cilj ovog Zakona je da poboljša funkcionisanje unutrašnjeg tržišta i promoviše primenu veštačke inteligencije koja je usmerena na ljude i kojoj se može verovati, uz obezbeđivanje visokog nivoa zaštite zdravlja, bezbednosti, osnovnih prava zagarantovanih *Poveljom o osnovnim pravima Evropske unije*, uključujući demokratiju, vladavinu prava i zaštitu životne sredine, od štetnih uticaja sistema veštačke inteligencije u Uniji, kao i podršku inovacijama. Oslanjajući se na sistem različitih nivoa rizika po društvo i pojedinca ova uredba predviđa odgovarajuće mere nadzora - uspostavlja uskladena pravila za tržište sistema veštačke inteligencije u Evropskoj uniji, uključujući zabrane određenih praksi, specifične zahteve za visokorizične sisteme i obaveze za operatere, pravila o transparentnosti i pravila za praćenje, nadzor i upravljanje tržištem, te propisuje mере за podršku inovacijama, sa fokusom na mala i srednja preduzeća i startapove ([16](#)).

Zakon o veštačkoj inteligenciji Evropske unije reguliše i generativne sisteme veštačke inteligencije, uključujući i aplikacije koje se koriste za kreiranje dipfejk sadržaja. On obavezuje kreatore da treniraju, dizajniraju i razvijaju generativni sistem veštačke inteligencije na način koji obezbeđuje da generisani sadržaji budu u skladu sa zakonima Evropske unije, kao i da dokumentuju i obezbede javno dostupan detaljan rezime o korišćenju podataka sa autorskim pravima za obuku sistema - a sve kako bi se ostvarila odgovarajuća zaštita od kršenja prava intelektualne svojine, posebno od kršenja autorskih prava. Pored toga kreatori generativnih modela se obavezuju na pridržavanje strožih obaveza u pogledu transparentnosti kako bi se izbeglo korišćenje generativnih sistema veštačke inteligencije za kreiranje manipulativnog sadržaja. U slučajevima kada je generativni sistem veštačke inteligencije korišćen za kreiranje dipfejk sadržaja (npr. teksta, videa ili zvuka koji deluje autentično ili istinito), korisnici koji su stvorili takav sadržaj moraju jasno naznačiti da je sadržaj generisan ili manipulisani veštačkom inteligencijom i kada je to moguće navesti ime pravnog ili fizičkog lica koje je generisalo ili manipulisalo sadržajem ([17](#)).

Stručnjaci iz ove oblasti ukazuju da je osnovna slabost ovakvog regulatornog pristupa nedostatak normativnih pragova. Dokumentovanje i javno objavljivanje može biti korisno samo ako postoje pragovi ili standardi za dobru ili lošu performansu kojih se kreatori generativnih sistema veštačke inteligencije moraju pridržavati. Isto važi i za korisnike generativne veštačke inteligencije. Kada bi postojali normativni pragovi, regulatorno opterećenje bi se prebacilo sa pojedinaca i nadzornih tela na pružaoce i korisnike, koji bi morali da kreiraju i implementiraju samo one modele i sisteme koji očigledno ispunjavaju relevantne normative zasnovane na riziku od kojih zakon polazi. Sandra Vačter slikovito objašnjava problematičnost ovog pristupa na sledećem primeru: Zamislimo bunar koji obezbeđuje vodu za neko selo, ali taj bunar sadrži visoke nivoe olova. Način na koji *Zakon o veštačkoj inteligenciji* pristupa generativnoj veštačkoj inteligenciji, umesto da rešava problem na izvoru, stanovnike sela samo obaveštava o toksičnosti vode, a na njima bi bilo da odluče da li će instalirati filtere za prečišćavanje. Efikasnije rešenje bilo bi uspostavljanje jasnih standarda za dozvoljene i nedozvoljene nivoe toksičnosti, uz obavezu preduzimanja preventivnih mera na samom izvoru. Vačter zaključuje da transparentnost o postojanju štetnih posledica nije isto što i preuzimanje odgovornosti za njih ([18](#)).

Etički izazovi generativne veštačke inteligencije

U stručnoj literaturi koja se bavi etikom generativne veštačke inteligencije najčešće se pominju pitanja odgovornosti, privatnosti i bezbednosti, problem pristrasnosti

generativne veštačke inteligencije, zatim problem autentičnosti sadržaja generisanih veštačkom inteligencijom i sa njom povezana pitanja atribucije, prava intelektualne svojine i autorskih prava, problem tačnosti i istinitosti generisanih sadržaja iz kojih proizilaze problemi zloupotrebe generativne veštačke inteligencije i sa njom povezano pitanja zaštite dostojanstva ličnosti. Pored toga autori iz ove oblasti bave se i problemom uticaja veštačke inteligencije na životnu sredinu ([19](#), [20](#), [21](#)).

Pristrasnost

Pristrasnost koja se vezuje za generativnu veštačku inteligenciju nije novi problem, već se nadovezuje na postojeće probleme detektovane u okvirima mašinskog učenja i razvoja algoritamskih sistema. Naime, generativni modeli se treniraju na postojećim setovima podataka iz kojih apsorbuju postojeće društvene predrasude. Na primer, rezultati istraživanja iz 2023. godine pokazuju da su dva testirana modela za generisanje slika (*DALL-E 2* i *Stable Diffusion*) na zahtev da generišu slike eksperata iz oblasti tehničkih nauka u 70 do 100% slučajeva isporučile slike na kojima su prikazani muškarci. Tako ovi modeli održavaju stereotipe o tehničkim naukama kao profesijama u kojima dominiraju muškarci – bilo da ih većinski privlače ili da u njima nadmašuju žene. U stvarnosti, zavisno od regionalnih i državnih razlika, žene trenutno čine između 28% i 40% diplomaca u oblasti tehničkih nauka ([22](#)).

U malom eksperimentu sprovedenom za potrebe ovog pregleda, alat za generisanje slika u okviru Čet Dži-Pi-Tija, na zahtev da generiše sliku novinara (pri čemu je kao unos korišćena rodno neutralna imenica na engleskom jeziku "journalist"), proizveo je dve slike koje prikazuju muškarce (slika 1), iako prema dostupnim podacima žene čine oko 40% predstavnika ove profesije ([23](#)).

journalist

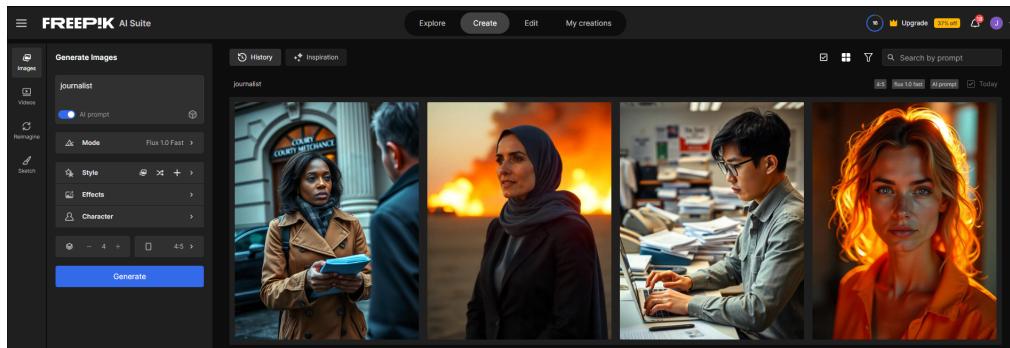


Here are the images depicting a journalist in action, capturing news in a bustling city environment. Let me know if there's anything else you'd like to see!



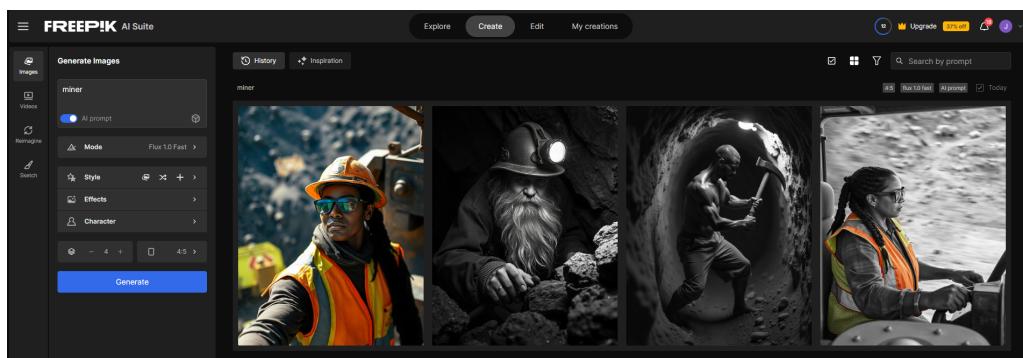
Slika 1. Skrinšot: AI Image Generator: journalist

Sa druge strane, stiče se utisak da su neki modeli u pokušaju da prevaziđu ovaj problem otišli u suprotnu krajnost. Na isti zahtev Fripik (*Freepik*) odgovara sa četiri slike, od kojih su na tri prikazane novinarke (slika 2).



Slika 2. Skrinšot: Freepik: Journalist

Isto čini i na instrukciju da generiše slike rudara (*miner*). Od četiri generisane, dve su slike žena (slika 3), mada prema podacima Svetske banke žene čine svega 15% radne snage u ovoj industriji (24). Međutim Fripik ovde podržava drugu vrstu stereotipa – rasne stereotipe, pa tako tri od četiri generisane slike predstavljaju crnce.



Slika 3. Skrinšot: Freepik: Miner

Pored pristrasnosti koja proizilazi iz pristrasnosti podataka na kojima se treniraju generativni modeli, pristrasnost može da bude utkana i u sam dizajn modela. To se pre svega odnosi na način kojim se određuju dizajnerski ciljevi koji definišu svrhu i funkcionalnost sistema, na pretpostavke dizajnera o tome ko će biti krajnji korisnici, zatim može da bude uslovljena kontekstom za koji je dati model dizajniran ili da proizlazi iz neadekvatne metodologije za testiranje i evaluaciju određenog modela (25).

Autentičnost, atribucija, prava intelektualne svojine i autorska prava

Čet Dži-Pi-Ti u veoma kratkom vremenu može produkovati sasvim pristojnu priču na bilo koju temu (slika 4), ili „umetničku“ sliku (slika 5).

ChatGPT ▾

napiši mi kratku priču o zimi

Zimska carolija

Sneg je tih prekrivao selo, pretvarajući ga u bajkoviti prizor. Mala Anja pritiskala je nos uz prozor, očiju punih sijaja. Napolju su se svetlucale pahulje, a vetar je nežno šaputao pesmu zime.

Jedva je dočekala da izade napolje. Navukla je rukavice i potrčala ka dvorištu, gde su njeni prijatelji već pravili sneška. Smeđi i vesela graja ispunili su hladan vazduh.

Dok su se grudvali, iznenada se začuo topao glas bake:

„Dodata, vruća čokolada je gotova!“

Deca su pojurala unutra, crvenih obraza i promrzlih ruku. U toplini bakinog doma, uz miris cimeta i vanile, zima je postala još čarobnija.

Ask anything



Search



Reason



Slika 4. Skrinšot: Čet Dži-Pi-Ti: Priča o zimi

U konkretnim slučajevima veštačka inteligencija je bez ljudske intervencije kreirala nova dela, istovremeno se oslanjajući na ljudske tvorevine koje se nalaze u bazama na kojima je trenirana. S tim u vezi postavlja se pitanje autentičnosti proizvoda, odnosno da li se ovaj čin generisanja sadržaja može izjednačiti sa ljudskim stvaralaštvom koje potiče iz čovekova istinskih iskustava i osećanja ([26](#)). Na ovo pitanje se nadovezuje i sledeće – pitanje atribucije. Posmatrano iz okvira etike, da li sadržaj generisan veštačkom inteligencijom treba pripisati korisniku, programeru modela ili nekoj drugoj strani. Ako isto pitanje postavimo u okvirima važećih zakona, ko je u tom slučaju nosilac autorskih prava? Za ovu komplikovanu situaciju za sada nema jednoznačnih rešenja. Iako su zakoni o zaštiti intelektualne svojine postavljeni tako da štite ljudsku kreativnost i inovativnost, što znači sadržaji generisani veštačkom inteligencijom ne bi trebalo da podležu zaštiti autorskih prava ([27](#), [28](#)), britansko, novozelandsko, indijsko i irsko zakonodavstvo omogućavaju „ograničena“ autorska prava za osobu koja je preuzela „potrebne aranžmane za stvaranje dela“ ([29](#)) ali ne i za kreatore sistema ili sam sistem veštačke inteligencije.

napravi mi sliku bele orhideje, ulje na platnu u Van Gogovom stilu



Evo slike bele orhideje u Van Gogovom stilu, sa dinamičnim potezima četkice i bogatom teksturom ulja na platnu. Nadam se da ti se dopada! ☺

[🔗](#) [🔗](#) [🔗](#) [🔗](#) [🔗](#) [🔗](#)

Slika 5. Skrinšot: Čet Dži-Pi-Ti: Orhideja

Osvrнимо se na primer slike orhideje. Čet Dži-Pi-Ti je generisao novu sliku u Van Gogovom stilu. Da li je etično da generativna veštačka inteligencija kreira sadržaj koji je sličan drugim umetničkim delima i tako zamagljuje granice između ljudske kreativnosti i potencijala za obmanu? Ovo pitanje reflektuju se i na pravnu zaštitu nosilaca autorskih prava čija su dela bila korišćena za treniranje modela veštačke inteligencije. Sposobnost veštačke inteligencije da stvara nove tekstove i slike koji su slični delima autora i umetnika čija su dela korišćena za obuku generativnih modela ili praksa da koristi celine ekstrahovane iz dela zaštićenih autorskim pravima u fokusu je nekoliko tekućih sudskih sporova. Na primer Džon Grišam, Džonatan Franc, Elin Hilderbrand ali i Njujork tajms tužili su Open AI zbog nelegalnog korišćenja njihovih dela i tekstova prilikom treniranje Čet Dži-Pi-Tija ([30](#)). Oni koji staju na stranu Open AI-a tvrde da bi obuka modela trebalo da se smatra fer upotreborom, jer mašinsko učenje predstavlja transformativnu upotrebu osnovnih podataka kao i da je licenciranje materijala za tako velike skupove podataka nepraktično zbog njihovog obima. Među braniocima Open-AI-a čuje se i argument da pitanje povrede autorskih prava nije primenjivo u ovom slučaju jer generativna veštačka inteligencija jednostavno nije zasnovana na kopiranju postojećih dela, već na stvaranju novih ([31](#)).

Tačnost i istinitost

Značajan problem koji se vezuje za generativnu veštačku inteligenciju jeste tačnost i istinitost sadržaja koje proizvodi. Naime tekstualni i vizuelni generativni modeli skloni su tome da proizvode lažne ili obmanjujuće rezultate - sadržaje koji mogu da zvuče ili izgledaju verodostojno, ali zapravo ne odražavaju činjenice. Ovaj fenomen poznat i kao halucinacije proizilazi iz prirode podataka za obuku koji mogu biti netačni,

nepotpuni ili zastareli, statističke prirode modela koja se zasniva na prepoznavanju i predviđanju obrazaca, kao i fokusa ovih modela da generišu odgovore koji bi trebalo da izgledaju verodostojno (32). Ovo može imate značajne posledice na upotrebu modela generativne veštačke inteligencije u obrazovanju, informisanju, medicini, ali može da ima i šire društvene posledice. U medijskoj žiji se krajem 2024. godine našla aplikacija Visper koja se širom sveta koristi za prevođenje i transkripciju intervjuja, generisanje tekstova u popularnim potrošačkim tehnologijama i kreiranje titlova za video snimke. Međutim ovaj alat veoma često halucinira - u transkripte ubacuje neizgovorene delove ili čak cele neizgovorene rečenice, uključujući rasne komentare, nasilnu retoriku i izmišljene medicinske tretmane. Kako se osim u bolnicama, trgovini i medijima na ovu aplikaciju u svakodnevnom životu oslanjaju i glavi i nagluvi oni predstavljaju posebno ranjivu kategoriju jer ne mogu znati da li je transkript tačan ili nije, što u krajnjem ovu ranjivu kategoriju stavlja u još nepovoljniji položaj u društvu (33).

Problem tačnosti sadržaja generisanih veštačkom inteligencijom ne proizilazi samo iz ograničenja modela već može biti i posledica zloupotrebe tehnologije. Kako bi se bolje razumele potencijalne opasnosti povezane sa zloupotrebama generativnih modela veštačke inteligencije Ferera ih slikovito prikazuje na sledeći način:

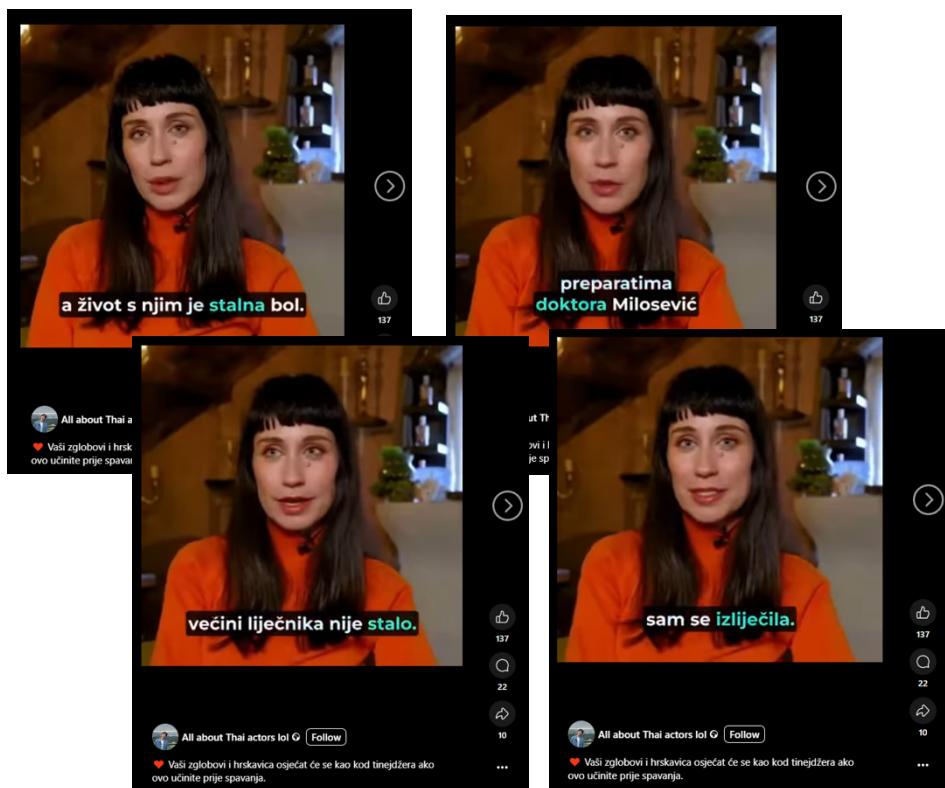
NAMERA				
Nepoštenje	Ciljano uznemiravanje	Manipulacije tržistem	Informacioni poremećaji	Ciljani nadzor
Propaganda	Digitalna imitacija	Ekstremističke sheme	Kampanje uticaja	Sintetičke realnosti
Obmana	Sintetički identiteti	Personalizovani otkup	Kontrola informacija	Sistemske aberacije (devijacije ili anomalije)
Vrsta štete	Lični gubitak i krada identiteta	Finansijska i ekonomski šteta	Manipulisanje informacijama	Društvena, socio-tehnička i infrastrukturna

Slika 5. Zlonamerna primena generativne veštačke inteligencije i velikih jezičkih modela, prilagođeno prema [Ferrara, 2024](#)

Ose matrice predstavljaju namere zlonamernih aktera i vrste štete koja može biti naneta. *Namere* zlonamernih aktera obuhvataju *Obmanu* (obmanjivanje radi lične koristi), *Propagandu* (promocija političkih, ideoloških ili komercijalnih ciljeva), i *Nepoštenje* (skrivanje ili iskrivljavanje istine za ličnu korist). Sa druge strane *Vrste štete* uzroko-

vane zlonamernim korišćenjem generativne veštačke inteligencije uključuju: „Štetu za osobu“ (kao što su krađa identiteta i povrede privatnosti), „Finansijsku i ekonomsku štetu“ (finansijske gubitke i prevare), „Manipulaciju informacijama“ (širenje dezinformacija i lažnih vesti) i „Društvenu, sociotehničku i infrastrukturnu štetu“ (šira šteta koja može uticati na zajednice, društvene strukture i ključnu infrastrukturu, uključujući pretnje demokratskim procesima, socijalnoj koheziji i tehnološkim sistemima).

Svaka ćelija predstavlja njihovu jedinstvenu kombinaciju i ilustruje višestruke oblike zloupotrebe generativne veštačke inteligencije. Na primer na preseku *propagande i manipulisanja informacijama* nalaze se kampanje uticaja. Jednu od takvih je tim FakeNews Tragača registrovao sredinom 2024. godine na društvenoj mreži Fejsbuk. Naime ovom mrežom se širio video-snimak u kojem pevačica Ana Đurić Konstrakta govori o tome kako je godinama patila od bolova u kolenima uzrokovanim osteoartritisom i kako je, nakon niza bezuspešnih pokušaja lečenja, konačno pronašla lekara i lek kojim se „potpuno izlečila“.



Ono što je privuklo pažnju redakcije [FakeNews Tragača](#) jeste činjenica da je ova Konstraktina ispovest generisana uz pomoć veštačke inteligencije (34). U konkretnom slučaju digitalne tehnologije su upotrebљene kako bi se izmenio snimak intervjuja koji je popularna pevačica dala zagrebačkom portalu [24.sata](#) u kojem nije pomenula nijedan

svoj zdravstveni problem (35) - a na pitanje da li veruje lekarima odgovorila je potvrđeno. Video je izmenjen u cilju promocije navodnog „pravog leka“ jednoj ranjivoj grupi ljudi bez pristanka Ane Đurić Konstrukte. Slične primere u više navrata detektovao je i fektčking portal Raskrinkavanje.ba, uključujući i dipfejk snimak (36) u kojem urolog i seksolog Aleksandar Milošević navodno govori da su problemi kao erektilna disfunkcija izazvani „kontaminacijom testosterona“ i da postoji „prirodna metoda“ za rešavanje takvih problema.

Na istom preseku *propagande i manipulisanja informacijama* nalazi se i primer videa generisanog veštačkom inteligencijom čiji je glavni akter bio francuski predsednik Emanuel Makron, a u kojem on navodno ljubi drugog muškarca (37, 38). Međutim, iako se u osnovi pozicioniraju na istom mestu u Ferarinoj matrici, cilj ove propagandne kampanje je narušavanje reputacije francuskog predsednika što u krajnjem treba da umanji nivo poverenja u ovog političara, kako u nacionalnom tako i u globalnom kontekstu.

Međutim iako dipfejk sadržaji sve češće postaju oružje u političkoj areni (39) i dalje se generativni modeli najčešće koriste za kreiranje seksualnih video zapisa žena bez njihovog pristanka. Prema procenama za 2023. godinu, dipfejk pornografija čini 98% svih dipfejk onlajn videa, od kojih 99% prikazuje žene. Na ovaj način javlja se jedan novi oblik rodno zasnovanog nasilja koji koristi veštačku inteligenciju za eksploraciju, ponižavanje i uzneniranje žena. Kao najčešće žrtve dipfejk pornografije i dalje se identifikuju poznate ličnosti zbog velikog broja slobodno dostupnih snimaka i fotografija uz pomoć kojih veštačka inteligencija generiše nove sadržaje (40). Međutim, zahvaljujući razvoju ovih tehnologija kao žrtve sve češće postaju i anonimne žene. U našem regionu je poslednjih par godina zabeleženo više slučajeva u kojima su žrtve bile maloletnice. Uz svega nekoliko snimaka sa društvenih mreža učenici dve osnovne škole na Novom Beogradu, generisali su pornografske video snimke i slike svojih školskih drugarica i nastavnica (41). Slični slučajevi zabeleženi su u Crnoj Gori, Severnoj Makedoniji, Rumuniji, Albaniji. Samo mali broj ovih slučajeva bude prijavljen, a onda se procesuiraju kao ugrožavanje privatnosti, krađa identiteta ili zloupotreba podataka (42).

O tome kakve su posledice po žrtve, možda najbolje govori svedočenje njihovih vršnjakinja: „Znam jednu devojčicu koja je htela da uradi ono najgore, ali se onda prenestila u drugi grad, jer nije mogla da podnese pritisak koji je trpela od vršnjaka, porodiće, prijatelja“.

Britanska pesnikinja Helen Mort, koja je takođe jedna od žrtava generisanih dipfejk slika, od kojih su neke prikazivale nasilan seks, u nekoliko navrata govorila je o svojim osećanjima i reakcijama koje su usledile po otkrivanju slika. U intervjuu za Gardijan kaže:

Želela sam – trebala mi je – mogućnost da razgovaram sa prijateljima i porodicom, ali sam se istovremeno osećala duboko posramljeno, kao da sam nekako sama kriva za to što mi se dogodilo. (...) Brinula sam se da će moji roditelji i stariji rođaci pomisliti da sam sama delila eksplicitne fotografije, da im koncept digitalne manipulacije neće biti poznat. Ali istovremeno sam bila besna zbog kulture koja podržava takve reakcije – okruženja u kojem se naglasak stavlja na to šta žrtve mogu da urade kako bi se zaštitile od zloupotrebe, umesto na to kako da sprecimo ljude da zloupotrebjavaju druge. Čak i da sam slala nage fotografije, to nikome ne bi dalo pravo da ih podeli sa svetom. (...) Takođe je bilo teško objasniti kako su fotografije koje nisu „stvarne“ mogле imati tako snažan uticaj na mene. Ali šta zapravo znači „stvarna“ slika? Imala sam noćne more o seksualnim napadima, direktno izazvane tim slikama. I dalje mogu da zatvorim oči i vidim ih, kao da sam zaista bila deo svake od tih scena.

Rezultati istraživanja ukazuju da žrtve dipfejk pornografije trpe ozbiljan psihološki stres, uključujući osećaje poniženja, straha, bespomoćnosti, gubitka kontrole i poteškoće u razvijanju poverenja u druge. Kod njih mogu da se razviju i mentalni poremećaji u vidu anksioznosti, depresije, posttraumatskog stresnog poremećaja i suicidalnih misli. Mnoge od njih dodatno viktimizuje osoba koja je generisala pornografski sadržaj, terajući žrtvu na ispunjavanje dodatnih zahteva, seksualnih ili nekih drugih, dok generisani sadržaj koristi kao sredstvo kontrole nad žrtvom. U svedočenju devojčice iz gorenavedenog primera vidimo na delu i viktimizaciju od strane društva. Žrtve neretko svedoče o izopštavanju iz društva, narušenom poslovnom ugledu, ocrnjavanju karaktera, uprkos tome što ni na koji način nisu učestvovali u kreiranju sadržaja ([43](#), [44](#)).

Odgovornosti, privatnost i bezbednost

Kao što se iz prethodno izloženog može videti, iako se zagovara odgovoran odnos prema razvoju i primeni generativne veštačke inteligencije, još uvek izostaju jasne smernice i mehanizmi kojima se utvrđuje ko je odgovoran u situacijama kada generativna veštačka inteligencija produkuje netačne ili štetne sadržaje i ko je odgovoran za sam proizvedeni sadržaj. Da li je to programer, kompanija, korisnik ili čak sam sistem veštačke inteligencije?

Etičke smernice za razvoj, primenu i upotrebu pouzdane i odgovorne veštačke inteligencije na ovo pitanje odgovaraju da su ljudska logika i prosuđivanje ključni „jer su ljudi ti koji pišu algoritme, definišu uspeh ili neuspeh, pripremaju podatke i skupove podataka, treniraju modele, vrše evaluaciju, donose odluke o upotrebi sistema“. Odgovornost za razmatranje uticaja sistema u okruženju u kome će biti primjenjen bi trebalo da bude na kompanijama koje su investirale u njihov razvoj i svim licima koja su uključena u njihovo stvaranje ([45](#)).

Kako generativna veštačka inteligencija postaje sve prisutnija u različitim oblastima i pitanja privatnosti i bezbednosti postaju sve važnija. Pitanje privatnosti može se posmatrati iz više uglova. Podaci na kojima se treniraju modeli i dalje sadrže informacije o ličnosti kao što su ime, adresa, finansijski podaci, čak i zdravstvene kartone, koji se neretko koriste bez izričitog pristanka osoba kojima pripadaju. Osnovna zabrinutost je ko ima ili može imati pristup ovim podacima, kako tokom redovnog funkcionisanja modela tako i u slučaju neželjenih preuzimanja ovih podataka. Pitanja privatnosti odnose se na curenje podataka, nadzor, profilisanje i lažno predstavljanje, dok se sigurnosni problemi tiču napada, manipulacije, zavisnosti i narušavanja integriteta sistema (46).

Sa druge strane i korisnici odaju različite tipove ličnih podatka u interakciji sa različitim modelima. Prema rezultatima istraživanja koje je sprovedla kompanija Harmonik (*Harmonic*). Njihovi istraživači su analizirali desetine upita upućenih Čet -Dži-Pi-Tiju, Kopajlotu, Džeminiju, Klaudu, Perpleksitiju tokom poslednjeg kvartala 2024. godine. U većini slučajeva poslovni korisnici koristili su generativnu veštačku inteligenciju za sažimanje teksta, uređivanje blogova ili pisanje dokumentacije za kod. Međutim, iako je većina upita bila benigna, oko 8,5% njih sadržalo je osetljive informacije kao što su detaljni podaci o korisnicima usluga, podaci o zaposlenima (koji su uključivali podatke od visine primanja do potpunih dosjeva zaposlenih), pravne i finansijske podatke, bezbednosne podatke (na primer bezbednosne politike, izveštaje o incidentima, evidenciju korisničkog pristupa) i osetljive kodove uključujući izvore i pristupne kodove (47)

Životna sredina

Pored svih drugih problema, brzi razvoj i upotreba generativne veštačke inteligencije izaziva zabrinutost i u vezi sa životnom sredinom. Najčešći problem na koji se ukazuje je velika potrošnja električne energije. Veliki modeli treniraju se na strukturama ogromne računalne snage, koji troše velike količine energije. Tu se potrošnja ne zaustavlja, već se uvećava tokom njihove upotrebe i u procesima usavršavanja. Visoka potrošnja električne energije za sobom povlači i rast emisije CO₂ i opterećenja elektroenergetske mreže.

Osim električne energije, tokom razvoja, upotrebe i usavršavanja modela potrebna je i velika količina vode koja hlađi hardvere, a njena povećana potrošnja takođe utiče na lokalne ekosisteme. Za proizvodnju visokoperformansnih računara potrebne su velike količine resursa poput litijuma, kobalta, aluminijuma i drugih, a njihova eksploracija dodatno ugrožava životnu sredinu. Tu je i pitanje odlaganja tehnološkog otpada u svim fazama, i mnoga druga.

UNESKO-ve *Preporuke o etici veštačke inteligencije* preporučuje državama članicama da „procene i smanje ekološki uticaj sistema veštačke inteligencije i infra-

struktura podataka, te obezbede usklađenost svih aktera u oblasti veštačke inteligencije sa zakonodavstvom, politikama i praksama zaštite životne sredine". Sproveđenje preporuka u delo pokazalo se teškim zadatkom i na primeru pripreme i usvajanja *Zakona o veštačkoj inteligenciji EU*. Ovaj zakon u nekoliko svojih članova osigurava da se sistemi koji su razvijeni ili se primenjuju u Evropskoj uniji moraju pridržavati standarda za održivost životne sredine. Međutim način na koji se reguliše ova oblast po mišljenju stručnjaka ima brojne manjkavosti. Zamera mu se to što se oslanja na dobrovoljne standarde, umesto na definisane obaveze, da se uglavnom bavi potrošnjom energije, dok u velikoj meri zanemaruje druge ključne ekološke probleme poput upotrebe vode, eksploatacije minerala i elektronskog otpada ([48](#), [49](#), [50](#)).

Eтика i primena generativne veštačke inteligencije u medijima i informisanju

Veštačka inteligencija, posebno generativni modeli, sve više se integriše u novinarstvo. Rezultati istraživanja sprovedenih krajem 2023. i početkom 2024. godine ukazuju da oko 40% novinara u Srbiji i 62% novinara u Bosni i Hercegovini koriste veštačku inteligenciju u procesu kreiranja medijskog sadržaja ([51](#), [52](#)) i to najčešće za istraživanje i prevod tekstova, pisanje delova tekstova odnosno bekgraundova, davanje naslova i pravljenje ilustracija i fotografija za članke. Međutim upotreba veštačke inteligencije nije vezana isključivo za procese obuhvaćene ovim istraživanjima već obuhvata celokupan informativni ciklus, od proizvodnje do diseminacije medijskih sadržaja i otvara novi ugao gledanja na već poznate etičke zabrinutosti koje se vezuju za upotrebu veštačke inteligencije - transparentnost, odgovornost, pristrasnost, kao i kulturnu, etničku, socijalnu i druge raznolikosti.

Transparentnost koja je immanentna etičkom novinarstvu zagovara se kao suštinska vrednost koja bi trebala da obezbedi odgovornu primenu veštačke inteligencije u ovoj profesiji. Međunarodna organizacija Reporteri bez granica objavila je u novembru 2023. godine *Parisku povelju o veštačkoj inteligenciji i novinarstvu* ([53](#)) u kojoj se insistira na transparentnom obeležavanju sadržaja generisanih veštačkom inteligencijom. Istu odredbu sadrži i novousvojeni *Kodeks novinara i novinarki Srbije* ([54](#)).

Kada je u pitanju odgovornost u novinarstvu, mediji i novinari su odgovorni za sadržaje koje proizvode i metode koje koriste. Sa uvođenjem sistema veštačke inteligencije u proces proizvodnje medijskih sadržaja problem odgovornosti se usložnjava jer algoritmi mogu delovati kao „crne kutije” što onemogućava uvid u proces donošenja odluka i otežava utvrđivanje odgovornosti za greške ili pristrasnosti u izveštavanju, bez obzira na to što se ona deklarativno pripisuje mediju i novinaru.

Što se tiče pristrasnosti i raznolikosti i tu nailazimo na nove grupe problema. Nepristrasnost u novinarstvu podrazumeva predstavljanje svih relevantnih činjenica i različitih perspektiva na način koji omogućava publici da formira sopstveno mišljenje. Međutim sistemi veštačke inteligencije koji se koriste za preporuku i personalizaciju sadržaja mogu selektovati sadržaje sa ciljem postizanja maksimalnog angažovanja publike umesto da se oslanjaju na novinarske vrednosti poput raznolikosti i pluralizma. Na ovaj način stvaraju se „filter mehuri” koji izoluju članove publike i ograničavaju pristup raznovrsnim informacijama. *Pariska povelja o veštačkoj inteligenciji i novinarstvu* u vezi sa ovim problemom nalaže „da bi dizajn i upotreba sistema veštačke inteligencije za automatsku personalizaciju sadržaja i preporuke trebalo da budu vođeni novinarskom etikom”. Međutim ovi sistemi najčešće nisu pod kontrolom medija već pod kontrolom tehnoloških kompanija koje se vode sopstvenim finansijskim interesima.

Sagledavajući svu kompleksnost problema u ovoj oblasti, stručnjaci ukazuju da je za očuvanje etičkih vrednosti u novinarstvu uz integraciju veštačke inteligencije u profesionalno delovanje koja nam neminovno sledi, neophodno obezbediti da „ljudski nadzor”, odnosno ljudska procena, kreativnost i urednički nadzor ostanu ključni deo novinarskog procesa i ne budu istisnuti iz redakcija različitim automatizovanim procesima ([53](#)).

Ova publikacija nastala je u okviru projekta „AI Verify: Provera dezinformacija generisanih veštačkom inteligencijom“ koji se realizuje uz podršku #SustainMedia Programme. Ovaj program je kofinansiran sredstvima Evropske unije i Nemačkog Saveznog ministarstva za ekonomsku saradnju i razvoj. Za sadržaj publikacije isključivo je odgovoran FakeNews Tragač, a sadržaj ne odražava nužno stavove EU ili Nemačkog Saveznog ministarstva za ekonomsku saradnju i razvoj.



IMPLEMENTIRAJU

