UDK: 316.774: 316.647.5/.8
004.8
Stručni rad
26. II. 2024.

LANA PRLIĆ*

UNIŠTENJE REPUTACIJE ALATIMA UMJETNE INTELIGENCIJE – *DEEPFAKE*

SAŽETAK

Govor mržnje smatra se jednim od izazova suvremenoga društva i demokracije. Ne postoji univerzalna definicija govora mržnje, a time ni jedinstven mehanizam borbe protiv njega. Umjetna inteligencija svojim brzim razvojem zauzima i ovaj aspekt u društvu kroz generiranje teksta u *online* prostoru, točnije društvenim mrežama, koje su najveći poligon za širenje govora mržnje. Govor mržnje alat je za uništavanje reputacije. Na društvenim mrežama cilj mu je i sustavno uništavanje reputacije kroz jasno targetiranje žrtve i gotovo orkestriran napad na osobu, kompaniju ili skupinu ljudi. Sustavno uništavanje reputacije znači oslabljivanje mete tako da se stavi pod upitnik, ponajviše, moral ili neko stajalište. Alati umjetne inteligencije, kao što je *deepfake,* koriste se upravo radi uništavanja reputacije. Digitalna moć kroz umjetnu inteligenciju, točnije alat *deepfake*, omogućava širenje ovakve vrste sadržaja. To je alat koji je lako dostupan te je, u biti, krivotvoren sadržaj koji se stvara putem programiranja i velikoga broja podataka, koji koristi sav materijal i sadržaj dostupan o određenoj budućoj žrtvi kojoj reputacija želi biti uništena. Metodom analize sadržaja u radu je prikazana analiza uništavanja reputacije putem *deepfakea*, kroz primjere, te su definirani pojmovi umjetne inteligencije, reputacije, uništavanja reputacije kroz alate umjetne inteligencije kao dio govora mržnje na društvenim mrežama, s posebnim naglaskom na *deepfake*.

Ključne riječi: umjetna inteligencija, reputacija, *deepfake*, govor mržnje

* University of Mostar, Mostar, Bosnia and Herzegovina, lana.prlic@gmail.com

UVOD

Umjetna inteligencija ne može se strogo definirati kao dobra ili loša, nego ona ovisi o načinu korištenja i motivima koji stoje iza korištenja njezinih alata. Alati umjetne inteligencije svojim brzim razvojem prepliću se sa govorom mržnje čiji je najveći poligon upravo na društvenim mrežama. Govoru mržnje na društvenim mrežama cilj je i sustavno uništavanje reputacije kroz jasno targetiranje žrtve i gotovo orkestrirani napad na osobu, kompaniju ili skupinu ljudi. Sustavno uništavanje reputacije znači oslabljivanje mete tako da se stavi pod upitnik ponajviše moral ili neko stajalište. Alati umjetne inteligencije, kao što je *deepfake*, koriste se upravo radi uništavanja reputacije. Metodologija korištena u radu jest kvantitativna, iz sekundarnih izvora kroz analizu sadržaja. Strukturu rada čine uvod, drugo poglavlje s definiranjem umjetne inteligencije, treće poglavlje koje sadrži osnovne pojmove o krivotvorenu sadržaju (*deepfake*) i uništenju reputacije, njihovoj povezanosti kroz analizu teorijskih spoznaja te primjenom na primjere, četvrto donosi zaključak i na kraju je literatura. Cilj je istraživanja utvrditi utjecaj *deepfakea* kao alata umjetne inteligencije na uništenje reputacije. Svrha je istraživanja dati preporuke u suzbijanju i prepoznavanju *deepfakea* u kontekstu uništenja reputacije.

Početne su hipoteze:
H1: *Deepfake* sredstvo je za uništenje reputacije.
H2: Umjetna inteligencija alat je za suzbijanje govora mržnje.
H3: *Deepfake*s ugrožava ljudska prava.

Digitalna moć kroz umjetnu inteligenciju, točnije alat *deepfake*, omogućava širenje ovakve vrste sadržaja. Osnovni je cilj umjetne inteligencije generirati podatke koristeći računala i mašine kako bi se doseglo oponašanje ljudskoga uma radi rješavanja problema. Algoritmi su gradivni elementi umjetne inteligencije. Štoviše, upravo zbog algoritma, koji je ključan element umjetne inteligencije, rezultati koji su proizvod umjetne inteligencije moraju se provjeriti, kritizirati i osporavati. Time dolazimo do odgovorne umjetne inteligencije, koja je postala tema mnogih inicijativa i projekata, a sve u svrhu suzbijanja štetnih utjecaja umjetne inteligencije. Iako se umjetna inteligencija može koristiti kao alat za suzbijanje govora mržnje u *online* prostoru, ovaj rad osvrnut će se konkretno na alat umjetne inteligencije naziva *deepfake*, koji u *online* prostoru dodatno širi govor mržnje, te sami pristup ovomu alatu. No, ni korištenje umjetne inteligencije u suzbijanju govora mržnje nije najbolje rješenje, s obzirom ma to da algoritmi ne prepoznaju kontekst, a dodatno otežava i činjenica da govor mržnje nema univerzalnu definiciju, što u jednu ruku dodatno otežava borbu protiv njega, na način koji se može nazvati suzbijanje slobode govora. *Deepfake* sintetički je medij u kojemu izgleda da netko nešto radi ili govori što netko nije uradio ili rekao. Ovako uvjerljivu sadržaju koji je krivotvoren pomoću računarskih programa, što postaje kasnije i viralno putem društvenih mreža, cilj je diskreditiranje targetirane osobe ili uništenje reputacije. *Deepfake* je alat za manipulaciju koji se može koristiti u razne svrhe, primjerice, osvetnička pornografija, manipulacija javnosti u političke svrhe, uništenje reputacije i slično. Utječe na *online* reputaciju tako što oštećuje kredibilitet osobe, nanosi štetu i može imati za posljedicu uništenje reputacije, gubitak posla te posljedice na mentalno zdravlje žrtve. Kao što je prethodno spomenuto, svjetske zvijezde već su se suočile sa zlouporabom umjetne inteligencije, točnije *deepfakeom*, kao što su Tom Hanks, Gayle King i Morgan Freeman. No, i u regiji *deepfake* ima svoje žrtve, primjerice Halida Bešlića.

UMJETNA INTELIGENCIJA I ALATI UMJETNE INTELIGENCIJE

Umjetna inteligencija skup je tehnika čiji je cilj da pomoću stroja, tehnološkim napretkom, reproduciraju ili imitiraju kognitivne sposobnosti ljudskoga bića. Nju su, prije svega, osmislili pa programirali upravo ljudi (Kostić i Sinders, 2022: 3). Ima širok utjecaj na donošenje odluka koje su bitne za ljude, štoviše zbog brza načina života ljudi čitaju vijesti bez provjere o točnosti informacija te na osnovi toga kreiraju mišljenje, javno mijenje. Umjetna inteligencija funkcionira na način da neki uređaj prima podatke koji su već prikupljeni,

obrađuje ih i daje odgovore, a ponekad i oponaša već prethodne slične situacije, sve ovisno o količini prikupljenih podataka i obradi istih. Umjetna inteligencija smatra se jednom vrstom nove revolucije kojoj se predviđa u budućnosti velik razvoj i veća uporaba nego danas. Alati koje ljudi koriste gotovo svakodnevno jesu alati umjetne inteligencije, primjerice, internetska kupovina i oglašavanje, digitalni osobni asistenti (Siri, Bixy), softveri za automatsko prevođenje i titlovanje, automobili, ali tu su i *deepfake*, *ChatGPT* itd. (Europski parlament, 2023). U ovome radu fokus je na *deepfake-u* koji se smatra negativnom stranom umjetne inteligencije, s obzirom na to da alati umjetne inteligencije poboljšavaju život, čine ga lakšim (osobni asistent, primjerice, ili aplikacije koje se koriste u svrhu zdravlja). Umjetna inteligencija postala je izazov i globalna tema u političkim krugovima radi rješavanja i definiranja, a sve u domeni ljudskih prava, informacija te slobode govora. Digitalna moć nije samo ograničena na moć koju imaju tehnološke kompanije nego i moći djelovanja samih korisnika. Umjetna inteligencija sastavni je dio moći koju tehnologije omogućavaju, tako što kreiraju mišljenje i oblikuju kontekst koji ne mora nužno biti istinit ili točan. Razvojem tehnologije brzina širenja sadržaja ne uključuje provjeru informacija prije plasiranja istih u javnost. Uporabom krivotvorenoga sadržaja (*deepfake*), manipulativnih videouradaka ili slika kao alata umjetne inteligencije ugrožava *online* reputaciju (Valeria G., 2023).

DEEPFAKES I UNIŠTENJE REPUTACIJE

Deepfakes prevodi se na hrvatski jezik i definira kao uvjerljivi krivotvoreni sadržaj, u obliku videosadržaja ili audiosadržaja koji se kreira putem umjetne inteligencije, a stvara „alternativnu realnost" u kojemu izgleda je netko nešto rekao ili uradio, a nije (Kostić i Sinders, 2022: 17). Taj krivotvoreni sadržaj ima moć uvjeriti nekoga da je lažno istinito i time nanijeti štetu nekomu drugomu iz motiva kao što su uništenje reputacije, politička manipulacija, *cyberbullying* te osvetnička pornografija (Valeria G., 2023). *Deepfake* krivotvori identitete osoba i pravi uvjerljiv sadržaj. Iako postoje programi koji mogu detektirati uporabu *deepfakea*, to ne umanjuje kršenje privatnosti i moći djelovanja korisnika koji je žrtva uništene reputacije. Vijeće Europe i Vijeće ministara Vijeća Europe još su 2020. godine poduzeli prve korake za zaštitu identiteta te definirali potrebu za posebno reguliranje sustava umjetne inteligencije s aspekta ljudskih prava, demokracije i vladavine prava. Štoviše, 2021. godine Odbor ministara osnovao je *Ad hoc odbor za umjetnu inteligenciju*, koji je usvojio dokument *Potencijalni elementi pravnog okvira za umjetnu inteligenciju zasnovanog na standardima Vijeće Europe o ljudskim pravima, demokratiji i vladavini prava*. Nacrt *Akta o digitalnim uslugama* regulatorni je instrument na razini Europske unije. Rizici umjetne inteligencije nestabilni su, promjenjivi i mogu biti štetni, primjerice *deepfake*, koji, pored reputacije, ljudskih prava i identiteta, dovodi i u pitanje prava na sadržaj, primjerice glas poznatih voditelja, pjevača ili glumaca. Tako su se Tom Hanks i Gayle King našli u središtu javnosti nakon što je upravo *deepfake* upotrebljen nad njima. Uporaba *deepfakea*, lažnih videosadržaja ili fotografija dovodi u pitanje reputaciju osobe i ugrožava je putem društvenih mreža. Narušava privatnost i ugrožava osobne podatke, točnije identitet, i stvara prijetnje koje prelaze granice *online* prostora.

Daljnjim tehnološkim napretkom raste i prijetnja reputaciji pojedinca kao i prijetnja cijelomu digitalnom okružju u kontekstu informacija, a upravo je najveća prijetnja *deepfake* sadržaj. No, negativno djelovanje *deepfake* sadržaja ne ograničava se samo na uništenje reputacije nego i na širenje lažnih vijesti, gubitak povjerenja, manipulacija te izazov pravnomu i zakonskomu odgovoru. Na više načina stvara neistinu sliku izmjenom facijalne ekspresije, kreiranjem novom lica na postojeće tijelo, zamjenom lica te zamjenom glasa, točnije korištenjem tuđega glasa i/ili za sadržaj koji određena osoba nije rekla. U samome početku bilo je jasno da se radi o parodijama, no brzim razvojem alata *deepfake*, laganim upravljanjem te jeftinijim programima korištenje *deepfake* alata preraslo je parodiju i postalo izazov slobodi govora, prijetnja uništenju identiteta i reputacije te širenju netočnih informacija. *Deepfake* upotrebljava se sa svrhom da se kompromitira neka osoba, sustavno uništi reputacija, pa ide čak toli-

ko daleko da se koristi i na način da se manipulira i pornografskim materijalom, *deepnude*s, radi osvetničke pornografije i/ili u političke svrhe. Sve navedeno stvara problem i sudovima, gdje su ovakvi slučajevi postali predmeti tužbi, a problem je nastao u nedovoljnu zakonskom okviru, reguliranju (Pepper i dr., 2021). *Deepfake* ima i elemente korištenja intelektualnih i autorskih prava kao što je glas, primjerice, što izravno ugrožava prava javnih osoba, voditelja, spikera, pjevača i sl.

Uništenje reputacije kroz *deepfake* nije zakonski definirano i zbog toga se u sudskim praksama često poziva na kršenje ljudskih prava, autorskih prava, prava na privatnost, kao i kleveta. Prakse koje su nabrojane nisu definirane, nego služe kao argumenti i načini borbe protiv *deepfake* sadržaja. Šteta koja je nanesena nečijoj reputacijom alatom kao što je *deepfake*, krivotvoreni sadržaj, teško je popravljiva zato što je reputacija jedna od najvrjednijih karakteristika koju pojedinac ili korporacija ima, izaziva lojalnost, poštovanje, dio je branda, pa čak i emotivne povezanosti s publikom. Tehnike u sustavnu uništavanju reputacije razvojem društvenih mreža poprimaju druge oblike, a društvene mreže, posebice govor mržnje, postaju jedan od glavnih alata pri uništavanju reputacije, s posebnim naglaskom na anonimnost. Uništavanje reputacije proces je koji se sastoji od dvaju elemenata. Prvi je element proces napada na pojedinca, tvrtku ili instituciju, a drugi je rezultat prvoga procesa, što u konačnici ima za cilj uništavanje reputacije. Proces napada ne mora imati „uporište", odnosno ne mora biti dijelom neke debate (Samoilenko i dr., 2018: 441-445). Uništavanje reputacije kao proces najviše se usredotočuje na preispitivanje moralnosti „mete" ili slabosti te ne mora imati za cilj samo uništavanje reputacije, nego i potpuno uništenje postojanja (Jasper, 2021). U *Routledge Handbook of Character Assassination and Reputation Management* detaljno je opisan cijeli proces uništenja reputacija kroz povijest, pa sve do modernoga doba i tehnologije koja je unaprijedila ovaj proces te ga napravila lakšim, bržim, posebice zbog anonimnosti, te dostupnim svima (Samoilenko i dr., 2020). Ovaj *Handbook* definira pet stubova uništenja reputacije: napadač, meta, medijski kanal, javnost i kontekst. Ta vrsta sustavna uništavanja reputacije zastupljena je među političarima, *influencerima*, neistomišljenicima, štoviše jasno je usmjerena prema novinarima, osobama iz javnoga života koji ne dijele mišljenje većine, LGBT populaciji, ženama i sl. Kanal su komunikacije društvene mreže koje, nudeći anonimnost i slab filter za govor mržnje, postaju poligon za ovu vrstu sustavna uništenja te kreiraju kontekst koji je svrsishodan za određeni motiv. U podizanju svijesti o opasnosti *deepfakea* bitna je uloga i poznatih osoba koje su se i same našle na meti krivotvorenoga sadržaja, a koje ćemo kroz ovaj rad spomenuti. U 2023. godini pažnju javnosti izazvao je Tom Hanks koji se javno obratio i skrenuo pažnju javnosti na to da je njegovo lice i glas korišteno za promociju dentalnih preparata. Nakon toga Gayle King obratila se javnosti kako bi objasnila da je video lažan, a sadržavao je njezin lik i glas u promociji proizvoda. Oboje su istakli da se mora biti oprezan s videima umjetne inteligencije, koji su napravljeni tako da izgledaju vjerodostojno. Ovaj problem postao je glasan i javan posebice nakon štrajka glumaca u Sjedinjenim Američkim Državama, koji se boje da će umjetna inteligencija i *deepfake* koristiti njihove izglede i glasove bez njihova odobrenja. Tu je i opasnost od *ChatGPT*-a koji predstavlja opasnost za scenariste. Objava Toma Hanksa skupila je na društvenoj mreži *Instagram* 128 767 sviđanja, a objava Gayle King 4901 sviđanje i 525 komentara. Video pod nazivom *This is not Morgan Freeman* na platformi *YouTube* postao je viralan i koristi se kao jedan od primjera koliko *deepfake* ili krivotvoreni sadržaj može biti realističan, vjerodostojan, da gledajući video, korisnik ne pomisli da je riječ o krivotvorenu sadržaju. Ovaj je video na platformi *YouTube* objavljen 2021. godine i ima preko 1 300 000 pregleda i preko 1000 komentara. Autor ovoga videa snimio je i kako je napravljen spomenuti video te objasnio kako izgleda *deepfake* „iza kamere". *Deepfake* kao alat umjetne inteligencije nije samo korišten za uništenje reputacije ili pravljenje krivotvorenoga sadržaja na javnim ličnostima u Sjedinjenim Američkim Državama nego i u Bosni i Hercegovini. Naime, bosanskohercegovački pjevač Halid Bešlić bio je njegova žrtva, kada su se njegov glas i izgled iskoristili za reklamiranje proizvoda za liječenje dijabetesa.

ZAKLJUČAK

Umjetna inteligencije postala je neizostavan element života u 21. stoljeću. Njezini alati korisni su i pomažu pri svakodnevnim obvezama te se koriste i za dizajniranje aplikacija i načina kojima se poboljšava kvaliteta svakodnevnoga života, okoline i zdravlja. Tako je ona postala neizostavan dio pojedinih profesija u svim sferama društva. No, kao i svako otkriće, i alat i umjetna inteligencija imaju svoje negativne strane, a jedna je od njih i krivotvoreni sadržaj (*deepfake*). Porast broja korištenja i jednostavna uporaba alata za *deepfake* predstavlja zakonske, poslovne i etičke izazove. Zakonska rješenja i prakse sudova ne prate brzinu razvoja ovih alata i izazova. Istraživanjem literature i analizom sadržaja potvrđene su prva i treća hipoteza, a druga je djelomično potvrđena istraživanjem i primjerima poznatih osoba koje su bile meta *deepfakea,* kao što su Gayle King, Tom Hanks i Halid Bešlić. Prva hipoteza da je *deepfake* sredstvo za uništenje reputacije potvrđena je jer *deepfake* koristi glas i lik osoba bez njihova pristanka na autentičan način, te je jedan od alata ne samo umjetne inteligencije nego i govora mržnje. S obzirom na to da *deepfake* koristi glas i lik osoba bez njihova pristanka, time se potvrđuje i treća hipoteza koja kaže da *deepfake* ugrožava ljudska prava. Druga hipoteza djelomično je potvrđena, a glasi da je umjetna inteligencija alat za suzbijanje govora mržnje. Točno je da se ona koristi i u tu svrhu, ali s obzirom na to da radi na principu prikupljanja podataka i obrade istih, pitanje je koji su ulazni podatci koje umjetna inteligencija skuplja i prima i na osnovu kojih eliminira, primjerice, komentare na društvenim mrežama. No, pokazala se kao jedan od dobrih alata u borbi protiv govora mržnje na društvenim mrežama, ali ne i dovoljnim. Borba protiv korištenja *deepfakea* u navedene svrhe moguća je poznavanjem samoga funkcioniranja *deepfakea*, točnije razvojem alata koji prepoznaju što je krivotvoreno ili lažno, a što vjerodostojno i istinito. Podizanje svijesti o ovim alatima umjetne inteligencije kroz medije i obrazovanje važan je segment borbe, ali nepotpuno je bez suradnje tehnološke industrije, vlade i znanosti kako bi se napravio sveobuhvatan odgovor na izazov i prijetnju koju predstavlja krivotvoreni sadržaj. Također, ne smiju se izostaviti poznate osobe kao netko tko ima moć podizanja svijesti radi definiranja i upoznavanja šire mase o štetnosti *deepfakea*. Važno je napomenuti da se ne koristi samo za uništenje reputacije poznatih i slavnih nego i u cilju osvetničke pornografije, govora mržnje, a sve navedeno vrlo brzo postane viralno i nosi sa sobom posljedice ne samo za reputaciju i ugled nego i za mentalno zdravlje.

LITERATURA

G., Valeria (2023) *The Future of Deepfakes in Online Reputation*, *Internet Reputation*, https://www.internetreputation.com/the-future-of-deepfakes-online-reputation/ (20. 11. 2023.).

Jasper, J. M. (2021) „Review Of The Routledge Handbook Of Character Assassination And Reputation Management", *Journal of Applied Social Theory*, 1(3), https://doi.org/https://socialtheoryapplied.com/journal/jast/article/view/101/98.

Kostić, B., Sinders, C. (2022) *Odgovorna umjetna inteligencija: Pregled uticaja umjetne inteligencije na ljudska prava i perspektive medijske pismenosti u kontekstu Bosne i Hercegovine.* Vijeće Europe, str. 3–17.

Pepper, C. i dr. (2021) *Reputation Management and the Growing Threat of Deepfakes*, *Bloomberg Law*,: https://news.bloomberglaw.com/us-law-week/reputation-management-and-the-growing-threat-of-deepfakes (20. 11. 2023.).

Samoilenko, S., Keohane, J., Icks, M. (2018) „Character assassination", E. Shiraev (ur.) *The Global Encyclopaedia of Informality: Understanding Social and Cultural Complexity*. London: UCL Press, pp. 441–445.

Samoilenko, Sergei i dr. (2020) *Routledge Handbook of Character Assassination and Reputation Management*, *Routledge & CRC Press*. Routledge, https://www.routledge.com/Routledge-Handbook-of-Character-Assassination-and-Reputation-Management/Samoilenko-Icks-Keohane-Shiraev/p/book/9781032081779 (10. 11. 2023.).

Što je umjetna inteligencija i kako se upotrebljava? (2023) | *Vijesti* | *Europski parlament*, https://www.europarl.europa.eu/news/hr/headlines/society/20200827STO85804/sto-je-umjetna-inteligencija-i-kako-se-upotrebljava (20. 11. 2023).

DESTRUCTION OF REPUTATION WITH ARTIFICIAL INTELLIGENCE TOOLS – *DEEPFAKE*

ABSTRACT

Hate speech is considered as one of the main challenges of modern society and democracy. There is no universal definition of hate speech, as well as there is no unique mechanism to fight against hate speech. Artificial intelligence by its fast development is a part of this context through generating of text in online space, more specific social media, which are the biggest place to spread hate speech. Hate speech is a tool for character assassination. Goal of the hate speech on social media is systemic character assassination though clear targeting of the victim, and orchestrated attack on person, company or certain group of people. Systemic character assassination means to weaken the target in a way to question moral or some of the stand points. Deepfake is one of the tools of artificial intelligence for the goal of character assassination. Digital power through artificial intelligence, specifically tool of deepfakes let this content to be spread. Deepfake is a tool which is easily accessible, and it is forged content which is made by programming and by a big number of data, material and content which is accessible about victim which reputation will be destroyed. By method of analysis of the content and examples in this article will be shown the analysis of character assassination by deepfake, as well as defined concept of artificial intelligence, reputation, and character assassination through the tools of AI as part of the hate speech on social media.

Key words: artificial intelligence, reputation, deepfake, hate speech.