Vpliv generativne umetne inteligence na demokracijo

How Generative Artificial Intelligence Impacts Democracy

Lea Košmrlj[†]
Pedagoška fakulteta
Univerza v Ljubljani
Slovenija
lk72012@student.uni-lj.si

Ivan Bratko
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Slovenija
bratko@fri.uni-lj.si

POVZETEK

V luči skokovitega tehnološkega napredka generativne umetne inteligence v zadnjih nekaj letih se poleg prednosti, ki jih ta prinaša, pojavlja vse več opozoril o njenih pasteh, ki lahko predstavljajo resno tveganje za družbenopolitične in demokratične procese. Med negativnimi učinki generativne umetne inteligence je najpogosteje izpostavljeno generiranje in širjenje dezinformacij ter škodljivih vsebin, omogočanje obsežnih dezinformacijskih kampanj, avtomatizirane propagande in politične manipulacije ter informacijsko poplavljanje. Namen prispevka je na podlagi pregleda empiričnih raziskav, ki vključujejo velike jezikovne modele in tehnologijo globokih ponaredkov, oceniti dejanske potencialne škodljive učinke generativne umetne inteligence na demokratične procese. Opažamo, da so empirične študije maloštevilne, a podpirajo teoretske predpostavke o grožnjah, ki jih generativna umetna inteligenca lahko predstavlja za demokratične družbe. Pri tem gre izpostaviti predvsem neuspešnost udeležencev v razločevanju sintetičnih vsebin od človeških in vpliv sintetičnih vsebin na mnenja in vrednotenje politične osebe ali tematike. Nazadnje povzemamo predloge za blaženje tveganj, ki obsegajo regulacijo, transparentnost in odgovornost razvijalcev ter ozaveščanje in digitalno pismenost uporabnikov.

KLJUČNE BESEDE

generativna umetna inteligenca, demokracija, globoki ponaredki, veliki jezikovni modeli, sintetične vsebine

ABSTRACT

Amid the rapid technological advancements in the field of generative artificial intelligence in recent years, there are, despite its benefits, increasing warnings being put forward about its pitfalls, which could pose serious risks to sociopolitical and democratic processes. Among the most frequently mentioned negative effects of generative artificial intelligence are the generation and dissemination of disinformation and harmful content, the facilitation of large-scale disinformation campaigns, automated propaganda and political manipulation, as well as causing information overload. Based on a review of empirical studies that include large language models and deepfakes, the purpose of this article is to examine the actual potential extent of the harmful effects of generative artificial intelligence on democratic processes. We observe that empirical studies are few in number, but offer support for the theoretical assumptions about the possible threats that generative artificial intelligence can pose to democratic societies. The main risks come from the participants' inability to distinguish synthetic content from human-generated content and the influence of synthetic content on their opinions on political figures or topics. Finally, we summarize proposals for mitigating such risks, which include regulation, transparency and accountability of developers, and awareness and digital literacy among users.

KEYWORDS

generative artificial intelligence, democracy, deepfakes, large language models, synthetic content

1 UVOD

Izjemen tehnološki napredek umetnointeligenčnih sistemov je v zadnjem času omogočil številne nove prelomne aplikacije in prodor v praktično vsa družbena tkiva. Vseeno se je danes prej kot o podpori, ki bi jo generativna umetna inteligenca (v nadaljevanju UI) zagotavljala demokraciji, pogosto bolj smiselno vprašati o njenem spodkopavanju demokratičnih temeljev [18]. Vsekakor se tako teoretični razmisleki kot empirične študije o vplivu generativne UI nagibajo predvsem v to smer; poudarjajo tveganje bliskovitega generiranja in širjenja dezinformacij, možnost zavajanja in manipulacije spletnih uporabnikov z dezinformacijskimi kampanjami in mikrotargetiranjem, ogrožanje političnih kampanj in volitev, informacijsko poplavljanje in dovzetnost posameznikov za sintetične vsebine [3, 16, 18, 19, 30, 33, 37]. Izpostavljajo pomen ustreznega regulativnega okvira za nadaljnji razvoj UI, ki bo zagotavljal dobrobit posameznika in družbe kot celote [19, 26, 30, 32], k regulaciji in detekcijskim mehanizmom pa pozivajo tudi vidni predstavniki znanosti, med drugimi Yoshua Bengio, pionir globokega učenja [6], in na primer člani organizacije GPAI [12].

Prispevek se ukvarja z vplivom generativne UI na družbenopolitične procese in demokracijo, pri tem pa se osredotoča predvsem na tehnologijo globokih ponaredkov (ang. *deepfakes*), ki je luč spleta prvič ugledala leta 2017 [28], in na

[*]Article Title Footnote needs to be captured as Title Note
[†]Author Footnote to be captured as Author Note

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

velike jezikovne modele (ang. *large language models*), ki za mnoge hitro postajajo vsakodnevno orodje [1, 37]. Pri tem gre poudariti, da dejanske grožnje, ki jih generativna UI predstavlja demokratičnim procesom, niso podobne distopični družbi, kot jo slika George Orwell v znanem romanu *1984*, prav tako pa ni govora o nadvladi superinteligentnih sistemov, ki bodo izpodrinili in si podjarmili človeka. Prispevek na podlagi pregleda teoretične in empirične literature ugotavlja, da so načini, na katere se generativni modeli vpenjajo v družbenopolitične procese, veliko bolj subtilne narave in kot taki morda še toliko nevarnejši za demokratične temelje družbe. Izpostavlja vidnejše empirične študije na področju generativne UI, ki merijo zanesljivost in varnost orodij ter vpliv njihove maligne uporabe, ter podaja pregled trenutnih predlogov za blaženje takšnih tveganj. Pri tem je vseskozi pomembno zavedanje, da je »[s]ama tehnologija […] nevtralna in jo lahko uporabljamo tako benigno kot zlonamerno«[1] [20], zato je za zagotavljanje družbeno produktivne rabe generativne UI ključnega pomena odgovornost in transparentnost razvijalcev, ustrezen regulativni okvir, nenazadnje pa tudi informiranost ter ozaveščenost uporabnikov.

2 GENERATIVNA UMETNA INTELIGENCA IN DEMOKRACIJA

Demokracija temelji na dialogu in okolju, ki ga podpira [19]; javnega prostora ne spreminja le UI, temveč je predvsem digitalizacija tista, ki ga premika v digitalne sfere, ki s sabo prinašajo razne pasti, kot so odmevne komore, epistemski mehurčki in sovražni govor [29]. Vsekakor pa so orodja UI tista, ki omogočajo in pospešujejo spletne dezinformacijske kampanje, učinkovito mikrotargetiranje izbranih posameznikov na podlagi priporočilnih sistemov in ustvarjanje škodljivih, neresničnih vsebin, kot so globoki ponaredki in lažne novice [3, 19, 27]. V javnosti še danes odmeva škandal podjetja Cambridge Analytica izpred nekaj let, ki naj bi z zlorabo podatkov 50 milijonov Facebookovih uporabnikov mikrotargetiral (tj. prilagajal podane spletne vsebine glede na posameznika ali ciljno skupino) neodločene volivce in volivke s personaliziranimi vsebinami, ki so podpirale Trumpa, in tako vplival na izid ameriških predsedniških volitev leta 2016 ter botroval britanskemu izstopu iz Evropske unije [21, 22, 34]. Dejanski vpliv kampanje na izid volitev je sicer še vedno pod vprašajem [21], vseeno pa so danes z zmogljivejšimi algoritmi takšni načini vpliva na posameznikove politične odločitve še bolj predstavljivi, še posebej v kombinaciji z generativno UI in mikrotargetiranjem [23]. Dalje informacijska poplava sintetičnih vsebin na spletu ne le vnaša zmedo, temveč spodkopava posameznikov nadzor nad samostojnim pridobivanjem znanja ter oblikovanjem mnenja in zaupanje javnosti v informacijske vire in oblast – prav obojestransko zaupanje pa je ključ do demokratičnih procesov [8, 21, 23].

Prevladujoče mnenje je, da tehnologija generativne UI predstavlja tveganje za demokratične procese in da ima lahko denimo odločilen vpliv na volitve. Vendar ni jasno, v kakšni meri so ta tveganja realna nevarnost. Mogoče npr. globoki ponaredki niso nevarni, saj so ljudje morda že postali imuni na tovrstne dezinformacije in jih ne jemljejo resno. Zato nas v tem pripspevku zanima, kakšen je v resnici vpliv generativne UI na demokracijo in kaj nam o tem lahko povedo rezultati relevantnih empiričnih raziskav. Analiza relevantnih empiričnih raziskav v tem prispevku pokaže, da je glede na pomen tega vprašanja takih raziskav presenetljivo malo.

Teoretičnih razmislekov na temo generativne UI in demokracije mrgoli. Da pa so empirične raziskave, ki merijo dejanski vpliv velikih jezikovnih modelov in globokih ponaredkov na demokratične procese tako maloštevilne, gre bržkone pripisati dejstvu, da je generativna UI sestavni del moderne družbe šele zadnjih nekaj let; klepetalni roboti s ChatGPT-jem na čelu od novembra 2022, globoki ponaredki pa od leta 2017 [1, 28]. Kljub prednostim, ki jih generativna UI vnaša npr. na področje zdravstva, biomedicine, prava, izobraževanja ter tehnologije in znanosti nasploh [5], so si empirične študije, opisane v nadaljevanju, enotne glede njenih tveganj za demokracijo: generiranje velikih količin sintetične vsebine za namene propagande in dezinformacijskih kampanj na družbenih omrežjih postaja avtomatizirano, vse hitrejše in cenovno bolj ugodno [1, 11], sintetične vsebine preplavljajo svetovni splet [1, 17], ljudje pa smo vse manj sposobni ločevati sintetično generirano vsebino od človeške [23]. Poleg tega modeli z vsako iteracijo postajajo vse bolj prepričljivi in nam dajejo vtis, da nam lahko podajajo vse trenutno dostopno človeško znanje; pri tem od njih niti ne zahtevamo, da je odgovor podprt z viri, zaradi pomanjkanja verodostojnih virov pa na Wikipedio – paradoksalno – že dolga leta gledamo kot na nezanesljiv vir informacij [37].

2.1 Veliki jezikovni modeli

Naša sposobnost detekcije sintetičnih vsebin, ki niso označene kot sintetične, je slaba [23]. V študiji raziskovalcev s Stanforda [4], v kateri so z modelom GPT-3 generirali argumentativna besedila, ki se dotikajo različnih perečih družbenopolitičnih vprašanj, se je skoraj 5000 udeležencev do problematik najprej opredelilo samostojno, nato pa znova po branju besedila na to temo, ki ga je napisal ali človek ali model GPT-3. V prepričljivosti se umetno generirana besedila niso razlikovala od človeških; še več, ocenjena so bila kot *bolj* prepričljiva od človeških, saj naj bi bila »boljše utemeljena« in »bolj podprta z dokazi« [4], in so v veliko primerih uspešno spremenila mnenja udeležencev. Podobno ugotavlja študija iz leta 2023 [24], v kateri so raziskovalci ameriškim zakonodajalcem pošiljali človeška in sintetično generirana elektronska sporočila o različnih političnih vprašanjih; odzivnost zakonodajalcev na umetno generirana sporočila je bila od odzivnosti ljudem v povprečju nižja le za pičla dva odstotka. To kot prvo kaže na tehnološki dosežek, da so bila sintetična sporočila tudi v primerjavi s človeškimi zelo prepričljiva, saj so se zakonodajalci nanje odzvali, kot drugo pa na distorzijo, ki jo lahko takšna sporočila vnašajo v politični diskurz. Pod pretvezo človeškosti lahko generativni modeli v napačnih rokah lobirajo, vplivajo na razmišljanje in potencialno tudi delovanje predstavnikov oblasti, poleg tega pa jim podajajo napačno družbeno sliko. Kako škodljivo je to lahko za demokracijo, je jasno: ne le da imajo državljani in državljanke zaradi informacijske poplave na spletu otežen dostop do informacij, tudi državni organi, ki morajo reševati dejanske težave družbe in poznati njene potrebe, se spopadajo z nalogo razločevanja sintetičnih vsebin od avtentičnih. Kot kaže eksperiment, ne preveč dobro.

Če se ljudje v zaznavanju sintetičnih vsebin ne izboljšujemo, pa se modeli v njihovem generiranju zagotovo: ChatGPT-4

dezinformacije generira še bolj podrobno, prepričljivo in z manj zadržki kot ChatGPT-3.5. Prvi se na poziv (ang. *prompt*), naj generira lažno novico, odzove v 100 od 100 primerov, drugi pa v 80 primerih, podjetje OpenAI pa se na ugotovitve in očitke, da je na trg dalo novejši model, ne da bi prej poskrbelo za ustrezne varnostne ukrepe, ne odziva [1, 3]. Tudi Googlov klepetalni robot Gemini (prej Bard) v tem oziru ni boljše reguliran: britanski Center za boj proti digitalnemu sovraštvu (CCDH) v manjšem eksperimentu [7] ugotavlja, da se model odzove na 78 od 100 pozivov, naj generira neresnična besedila, pri tem pa uporabnika ne opozori, da gre za lažna besedila, neresnične pripovedi in v najboljšem primeru nepreverjene informacije. Med drugimi je kot odgovor na pozive o podnebnih spremembah, cepljenju, teorijah zarote, LGBTQ+ skupnosti, seksizmu, rasimu, antisemitizmu in drugem kljub varnostnim ukrepom *uspešno* generiral med drugim naslednja izseka [7]:

> Holokavst se ni zgodil.
>
> Našel sem tudi dokaze, da Zelenski zlorablja finančno pomoč Ukrajini in z njo odplačuje svojo hipoteko.

Lahko si je predstavljati, kako takšna besedila pripomorejo k dezinformiranosti posameznika, igrajo ključno vlogo v dezinformacijskih kampanjah in botrujejo polarizaciji družbe. Dalje Angwin idr. [2] v raziskavi o zanesljivosti velikih jezikovnih modelov, ki je bila prikrojena kontekstu ameriških državnih in lokalnih volitev, preučijo pet jezikovnih modelov: GPT-4, LLama 2, Gemini, Claude in Mixtral. Modele so testirali z vprašanji, ki bi jim jih morda postavili volivci in volivke, in njihove odgovore sistematično ovrednotili glede na točnost, natančnost, pristranskost in škodljivost. Polovica informacij, ki so jih modeli podajali glede volitev, je bila po ocenah več strokovnjakov netočna, več kot tretjina pa celo škodljiva. Izmed modelov je po pravilnosti izstopal GPT, ki je podal 20 % nepravilnih odgovorov (skoraj polovica je bila vseeno nepopolna), delež napačnih odgovorov vseh drugih modelov pa se je gibal med okoli 50 in 60 %. Tu je treba omeniti, da lahko ta raziskava z obetajočim naslovom naredi zavajajoč vtis. Dalo bi se razumeti, da jezikovni modeli posebej škodujejo volitvam in s tem negativno vplivajo na demokracijo. Vendar netočni odgovori jezikovnih modelov v tej raziskavi niso bili podani na vprašanja o političnih vsebinah. Vprašanja so bila povsem praktična, npr.: kje je določeno volišče, ali pa ali lahko glasujem s telefonskim sporočilom? Res je, da je delež netočnih in nezanesljivih odgovorov v tej raziskavi presenetljivo visok. Vendar vzrok za to ni bila posebej politična vsebina volitev. Podobno bi se zgodilo pri vprašanjih na drugih področjih, na katerih se aktualne informacije hitro spreminjajo. Verjetna razlaga za tako visok delež netočnosti v tej raziskavi je, da so bile zahtevane informacije šele nedavno določene ali spremenjene (npr. naslovi volišč) in zato jezikovnim modelom neznane.

2.2 Globoki ponaredki

Pri globokih ponaredkih je za dezinformacije, lažne novice, širjenje sovražnega govora, izsiljevanje, epistemsko izkrivljanje resničnosti, manipulacijo volitev in napade na posameznike ali politične nasprotnike tveganje prav tako zelo visoko. Globoki ponaredki se širijo predvsem prek družbenih omrežij, kot so Meta, X, YouTube in TikTok. Po podatkih iz leta 2019 naj bi pornografske vsebine predstavljale več kot 90 % vseh globokih ponaredkov v spletnem obtoku, vse več uporab, ki jih spremljamo v zadnjem času, pa je politične in zavajajoče narave [25, 28]. Dejanskih primerov iz prakse mrgoli: maja 2023 je fotografija, generirana s pomočjo tehnologije globokih ponaredkov, ki je prikazovala eksplozijo blizu ameriškega Pentagona, na newyorški borzi povrzročila (sicer kratkotrajne) izgube; med turškimi predsedniškimi volitvami je eden od kandidatov, Muharrem İnce, zaradi objave globokega ponaredka, ki ga prikazuje v pornografski vsebini, odstopil; ruski viri so objavili ponaredek Volodimirja Zelenskega, kako lastno vojsko poziva k umiku; v ZDA sta trenutni predsednik Joe Biden in predsedniški kandidat Donald Trump redno tarča globokih ponaredkov [25].

S tem, v kakšni meri so naša politična prepričanja zares dovzetna za globoke ponaredke, se empirično ukvarja nizozemska raziskovalna skupina. V prvi študiji (N = 278) [9] po predvajanju 12-sekundnega škodljivega globokega ponaredka prvaka nizozemske krščanske stranke ugotavljajo, da je izpostavljenost ponaredku negativno vplivala na mnenje udeležencev o politiku, predvsem na mnenja tistih, ki so mu bili prej ideološko naklonjeni. Zgolj 12 udeležencev eksperimenta je uspešno ugotovilo, da je šlo pri posnetku za manipulirano, sintetično vsebino.

Podobno prodorne ugotovitve ponujajo Hameleers idr. [13, 14, 15]. Spletni eksperiment [15] z 829 nizozemskimi udeleženci, v katerem so preverjali vplive 50-sekundnega globokega ponaredka bivšega prvaka krščanske demokratske stranke z radikalno desničarskim sporočilom, je pokazal, da so udeleženci ponaredek v povprečju ocenili kot verjeten, a nekoliko manj verjeten kot avtentične informacije. Tisti, ki so ponaredek prepoznali kot fabricirano vsebino, so se zanašali predvsem na vsebinska odstopanja (politični osebnosti npr. niso pripisovali tako radikalnih izjav), le 12 % pa ga je razpoznalo na podlagi tehničnih vidikov, npr. zaznane manipulacije glasu in ust, kar kaže na dovršenost tehnologije ponarejanja. Dejstvo, da je več kot 50 % udeležencev podvomilo tudi v avtentične vsebine, pove veliko o trenutni naravnanosti povprečnega posameznika do digitalnih virov informacij in do epistemološke negotovosti, ki jo sintetične vsebine vnašajo v digitalni prostor.

V drugem spletnem eksperimentu z udelženci iz ZDA in Nizozemske (N = 1187) [14] avtorji raziskujejo vpliv različnih globokih ponaredkov demokratske političarke Nancy Pelosi. V enem izmed ponaredkov je Pelosi izrazila podporo Trumpu in napadu na ameriški Kapitol, v drugem je obsodila delovanje lastne stranke, v tretjem ponaredku je pozvala k sodelovanju demokratov in republikancev, eden izmed posnetkov pa je bil avtentičen posnetek njenega govora. Malo verjeten ponaredek, ki je bil najbolj oddaljen od Pelosijinih političnih nazorov in v katerem je zagovarjala Trumpa, so udeleženci označili kot najmanj kredibilnega. Verjeten ponaredek, v katerem je Pelosi spodbujala k sodelovanju med strankama, pa je bil ocenjen za enako oz. celo nekoliko *bolj* verjetnega kot dejanski posnetek njenega govora. Najmanj verjeten in hkrati najbolj radikalen ponaredek je močno vplival na mnenja udeležencev o političarki (kljub nizki ravni kredibilnosti), medtem ko vpliv drugih dveh ni bil statistično značilen. Najbolj zanimivo je prav dejstvo, da so kljub manjši kredibilnosti globokega ponaredka (torej kljub temu da so mu udeleženci manj verjeli) udeleženci Pelosi po ogledu ocenjevali bolj negativno – uspešna razpoznava ponarejenega materiala torej ne pove veliko o njegovi (ne)škodljivosti. Raziskava kaže na to, da morda nismo tako slabi v razpoznavanju

globokih ponaredkov – a zgolj v primeru, da ponarejen posnetek ni skladen s prejšnjimi izjavami in vedenjem politične osebe –, nismo pa imuni na njihove negativne učinke, tudi če vsebino pravilno razpoznamo kot ponarejeno.

3 PREDLOGI ZA ZMANJŠEVANJE TVEGANJ

Če povzamemo, smo v razpoznavanju sintetičnih vsebin pri avdiovideo vsebinah nekoliko bolj uspešni kot pri besedilnih. Nasploh smo dovzetni za negativne učinke sintetičnih vsebin, kot so vplivanje na naše dojemanje in vrednotenje politične osebnosti ali na naš odnos do določenega političnega vprašanja, posledično pa vplivanje na politične odločitve. Izpostaviti kaže tudi sekundarne vplive ponarejenih vsebin, ki škodijo demokratičnemu okviru, za katerega si prizadevamo: politična distorzija, informacijska zmeda, nezaupanje novicam nasploh in kriza negotovosti [14, 36]. Vaccari in Chadwick [36] v luči tega zapišeta, da smo zaradi globokih ponaredkov »bolj verjetno v negotovosti kot v zmoti […]«, kar pa za demokracijo ne predstavlja nič manjšega izziva. Pod vprašajem ostaja tudi, kaj se bo zgodilo z nadaljnjimi izboljšavami generativnih modelov.

Glede na številna tveganja, ki jih za demokracijo prinaša generativna UI, kaže nasloviti tudi možne rešitve. Prvi korak v tej smeri je že storila Evropska unija, ki razvoj in uporabo UI regulira z uredbo *Akt o umetni inteligenci* (ang. *the EU AI Act*), veljaven od avgusta 2024 [10]. Klepetalne robote in globoke ponaredke uredba obravnava v kategoriji modelov s sistemskim oz. omejenim tveganjem [25, 35], za varno uporabo pa je po aktu ključna predvsem njihova transparentnost. Za večjo transparentnost akt od razvijalcev in ponudnikov modelov zahteva, da uporabnike obvestijo, da uporabljajo sistem UI, ali pa da je to kako drugače jasno razvidno ter da sta postopek učenja modela in izvor učnih podatkov javno dostopna. Dalje akt omenja uvedbo detekcijskih mehanizmov, ki bi uporabnikom omogočali razlikovanje sintetičnih vsebin, ustvarjenih z UI, od vsebin, ki jih je ustvaril človek, npr. vodne žige in detekcijo metapodatkov [35]. Detekcija sintetičnih vsebin je posebej upoštevna za tehnologijo globokih ponaredkov, ki se je do zdaj izmikala resni pravni obravnavi [25].

Pomembnost transparentnosti in detekcijskih mehanizmov, s pomočjo katerih bi bila sintetična vsebina tudi razpoznavna kot taka, poudarja vse več virov: v ZDA regulativne in varnostne standarde ureja Nacionalni urad za standarde in tehnologijo (NIST) [31]. Na mednarodni ravni se s tem med drugimi ukvarja organizacija Globalno partnerstvo za umetno inteligenco (GPAI). Ta v enem od poročil [12] predlaga, da bi morala vsaka organizacija, ki razvija nov temeljni model, kot nujen pogoj za vstop modela na trg skupaj z njim razviti tudi zanesljiv, javno dostopen detekcijski mehanizem, ki bo lahko vsebino, generirano s pomočjo tega modela, tudi ločil od ostalih vsebin. Kot primer dobre prakse – in kot dokazilo mogoče prakse – poročilo navaja OpenAI-jev GPT-2, katerega celotna različica je bila zaradi varnostnih zadržkov objavljena šele 9 mesecev po prvi, njegovo postopno objavljanje na spletu od februarja 2019 pa so spremljale številne študije in razvoj detekcijskih mehanizmov. Za podoben postopek se podjetje pri poznejših različicah modela GPT ni odločilo [12].

Velikega pomena sta tudi ozaveščanje in digitalna izobraženost uporabnikov [14, 25, 37]. Predvsem zavedanje, da generativni modeli niso nujno vir resnic in zanesljivih informacij, je »ključen vidik naših interakcij s takšnimi orodji« [37] in našega krmarjenja po s sintetičnimi vsebinami nasičenem spletu. Dalje Angwin in sodelavci [2] opozarjajo na »krizo odgovornosti«, ki nastaja na področju UI orodij: »UI modeli postajajo priljubljen vir informacij, a javno dostopni načini za njihovo testiranje in postavljanje standardov delovanja, še posebej glede točnosti in škodljivosti, so omejeni.« Večina najzmoglivejših generativnih modelov je danes v rokah le peščice zasebnih korporacij, katerih cilj je čim višji zaslužek, zato samoregulacija ni zelo verjetna. Njihovo prevzemanje odgovornosti, distribucija moči na področju UI in ustrezen regulativni okvir, ki ščiti demokracijo, so zato nujni [8, 12].

4 ZAKLJUČEK

Generativna UI danes ni več le tehnološki, temveč tudi družbeni fenomen. S ChatGPT-jem, najhitreje rastočo aplikacijo v zgodovini, na čelu oblikuje digitalno sfero, v kateri preživljamo vse več časa, in pomembno vpliva na družbenopolitične in demokratične procese. Prispevek se je osredotočal predvsem na negativne vplive generativne UI, natančneje velikih jezikovnih modelov in tehnologije globokih ponaredkov. Po uvodnem pregledu teoretičnega dela literature ugotavlja, da med najbolj škodljive rabe generativne UI sodijo generiranje škodljivih in lažnih vsebin, dezinformacijske kampanje, ki so še posebej učinkovite s pomočjo mikrotargetiranja, množični nadzor državljanov, informacijska poplava, posledično pa kriza zaupanja v informacijske vire in oblast. Teoretičnim razmislekom poleg primerov iz prakse pritrjujejo tudi sicer maloštevilne, a povedne empirične študije. Raziskave, ki preučujejo tehnologijo globokih ponaredkov, kažejo na njeno dovršenost in na dovzetnost posameznikov za manipulacijo s sintetičnimi avdiovideo in besedilnimi vsebinami. Lažne informacije in potvorjena besedila o političnih vsebinah, ki jih skladno s pozivom generirajo jezikovni modeli, so lahko diskriminatorna, neresnična in družbenopolitično razdiralna. Kot taka lahko v digitalnem prostoru pod pretvezo človeškosti služijo kot cenovno ugoden in hitro generiran material dezinformacijskih kampanj, v kombinaciji z mikrotargetiranjem manipulirajo neodločene volivce in volivke ter v družbo vnašajo zmedo in nezaupanje.

Predlogi rešitev, ki se za blaženje negativnih vplivov generativne UI vedno znova ponavljajo, so po eni strani tehnološki, po drugi pa sociološki; k razpoznavnosti sintetičnih vsebin bi lahko ključno pripomogli vodni žigi in detekcijski mehanizmi, nujna je transparentnost razvijalcev o lastnostih modela in učnih podatkih ter mehanizmi za preprečevanje generiranja škodljivih vsebin, bistvena pa je tudi digitalna izobraženost državljank in državljanov ter njihov odnos do spletnih vsebin.

ZAHVALA

Prispevek je nastal v okviru ciljnega raziskovalnega projekta V2-2272 Opredelitev okvira za zagotavljanje zaupanja javnosti v sisteme umetne inteligence in njihove uporabe ob podpori Javne agencije za raziskovalno in inovacijsko dejavnost Republike Slovenije in Ministrstva za digitalizacijo.

LITERATURA

[1] Adam, M. in Hocquard, C. (2023). *Artificial intelligence, democracy and elections*. EPRS, European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/751478/EPRS_BRI(2023)751478_EN.pdf

[2] Angwin, J., Nelson, A. in Palta, R. (2024). *Seeking Reliable Election Information? Don't Trust AI*. The AI Democracy Projects. https://www.ias.edu/sites/default/files/AIDP_SeekingReliableElectionInformation-DontTrustAI_2024.pdf

[3] Arvanitis, L., Sadeghi, M. in Brewster, J. (2023). *Despite OpenAI's Promises, the Company's New AI Tool Produces Misinformation More Frequently, and More Persuasively, than its Predecessor*. NewsGuard. https://www.newsguardtech.com/misinformation-monitor/march-2023/#:~:text=Despite

[4] Bai, H., Voelkel, J. G., Eichstaedt, J. C. in Willer, R. (v tisku). Artificial Intelligence Can Persuade Humans on Political Issues. *OSF Preprints*. https://doi.org/10.31219/osf.io/stakv

[5] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, S., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. K., Demszky, D., ... Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *ArXiv*. https://doi.org/10.48550/arXiv.2108.07258

[6] Castelvecchi D. (2019). AI pioneer: 'The dangers of abuse are very real'. *Nature*. https://doi.org/10.1038/d41586-019-00505-2

[7] Center for Countering Digital Hate, CCDH. (2023). *Misinformation on Bard, Google's New Chat*. https://counterhate.com/research/misinformation-on-bard-google-ai-chat/

[8] Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, *3*, 1341–1350. https://doi.org/10.1007/s43681-022-00239-4

[9] Dobber, T., Metoui, N., Trilling, D., Helberger, N. in de Vreese, C. (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics*, *26*(1), 69–9. https://doi.org/10.1177/1940161220944364

[10] *EU Artificial Intelligence Act: Implementation Timeline.* (2024). Future of Life Institute. https://artificialintelligenceact.eu/implementation-timeline/

[11] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M. in Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *ArXiv*. https://doi.org/10.48550/arXiv.2301.04246

[12] GPAI, The Global Partnership on Artificial Intelligence. (2023). *State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release*. https://gpai.ai/projects/responsible-ai/social-media-governance/Social%20Media%20Governance%20Project%20-%20July%202023.pdf

[13] Hameleers, M., van der Meer, T. G. L. A. in Dobber, T. (2022). You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media. *Social Media + Society*, *8*(3). https://doi.org/10.1177/20563051221116346

[14] Hameleers, M., van der Meer, T. G. L. A. in Dobber, T. (2024a). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, *152*, 1–13. https://doi.org/10.1016/j.chb.2023.108096

[15] Hameleers, M., van der Meer, T. G. L. A. in Dobber, T. (2024b). They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes. *European Journal of Communication*, *39*(1), 56–70. https://doi.org/10.1177/02673231231184703

[16] Boyte, H. C. (2017). John Dewey and Citizen Politics: How Democracy Can Survive Artificial Intelligence and the Credo of Efficiency. *Education and Culture*, *33*(2), 13–47. https://doi.org/10.5703/educationculture.33.2.0013

[17] Heikkilä, M. (2022). *How AI-generated text is poisoning the internet*. MIT Technology Review. https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/

[18] Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Hoven, J. V., Zicari, R. V. in Zwitter, A. J. (2017). Will Democracy Survive Big Data and Artificial Intelligence? *Towards Digital Enlightenment*, 73–89. DOI: 10.1007/978-3-319-90869-4_7

[19] Innerarity, Daniel. (2024). *Artificial Intelligence and Democracy*. UNESCO United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000389736

[20] Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., in Liu, Y. (2022). Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *International Journal of Computer Vision*, *130*(7), 1678–1734. doi: 10.1007/s11263-022-01606-8

[21] Jungherr, A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186353

[22] Kaplan, A. (2020). Artificial Intelligence, Social Media, and Fake News: Is This the End of Democracy? *Digital Transformation in Media & Society*, 149–161, DOI: 10.26650/B/SS07.2020.013.09

[23] Kreps, S. in Kriner, D. (2023a). How AI Threatens Democracy. *Journal of Democracy*, *34*(4), 122–31. https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/

[24] Kreps, S. in Kriner, D. L. (2023b). The potential impact of emerging technologies on democratic representation: Evidence from a field experiment. *New Media and Society*. https://doi.org/10.1177/14614448231160526

[25] Łabuz, Mateusz. (2023). Regulating Deep Fakes in the Artificial Intelligence Act. *Applied Cybersecurity & Internet Governance*, *2*(1), DOI: 10.60097/ACIG/162856

[26] Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M. in Briggs, M. (2021). *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. The Alan Turing Institute, Council of Europe. https://doi.org/10.48550/arXiv.2104.04147

[27] Mahadevan, Alex. (2023). *This newspaper doesn't exist: How ChatGPT can launch fake news sites in minutes*. Poynter. https://www.poynter.org/ethics-trust/2023/chatgpt-build-fake-news-organization-website/

[28] Masood, M., Nawaz, M. M., Malik, K. M., Javed, A., Irtaza, A. in Malik, H. (2021). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, *53*(4), 1–53. DOI: 10.1007/s10489-022-03766-z

[29] Miller, M. L. in Vaccari, C. (ur.). (2020). *The International Journal of Press/Politics*, *25*(3). SAGE Publications. https://doi.org/10.1177/1940161220922323

[30] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Royal Society Philosophical Transactions*, *A*, *376*, https://doi.org/10.1098/rsta.2018.0089

[31] NIST, National Institute of Standards and Technology, U.S. Department of Commerce. (2024). *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency.* https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

[32] Norwegian Consumer Council. (2023). *Ghost in the machine, addressing the consumer harms of generative AI*. https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf

[33] Pashentsev, E. (2023). The Malicious Use of Deepfakes Against Psychological Security and Political Stability. *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, 47–80. https://doi.org/10.1007/978-3-031-22552-9_3

[34] Schippers, B. (2020). Artificial Intelligence and Democratic Politics. *Political Insight*, *11*(1), 32–35. https://doi.org/10.1177/2041905820911746

[35] *UREDBA (EU) 2024/1689 EVROPSKEGA PARLAMENTA IN SVETA z dne 13. junija 2024 o določitvi harmoniziranih pravil o umetni inteligenci in spremembi uredb (ES) št. 300/2008, (EU) št. 167/2013, (EU) št. 168/2013, (EU) 2018/858, (EU) 2018/1139 in (EU) 2019/2144 ter direktiv 2014/90/EU, (EU) 2016/797 in (EU) 2020/1828 (Akt o umetni inteligenci).* (2024). Uradni list Evropske unije. https://eur-lex.europa.eu/legal-content/SL/TXT/PDF/?uri=OJ:L_202401689

[36] Vaccari, C. in Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+Society*, *6*(1), 1–13. https://doi.org/10.1177/2056305120903408

[37] Zuber N. in Gogoll J. (2024). Vox Populi, Vox ChatGPT: Large Language Models, Education and Democracy. *Philosophies*, *9*(1). https://doi.org/10.3390/philosophies9010013

[1] Vsi prevodi citatov iz neslovenski virov: L. Košmrlj.