VPLIV GENERATIVNE UMETNE INTELIGENCE NA DEMOKRATIČNE PROCESE

Lea Košmrlj in Ivan Bratko

Uvod[1]

Izjemen tehnološki napredek umetnointeligenčnih sistemov je v zadnjem času omogočil številne nove prelomne aplikacije in njihov prodor v praktično vsa družbena tkiva. Vseskozi pa razvoj generativne umetne inteligence (v nadaljevanju generativna UI) spremljajo tako teoretični razmisleki kot empirične študije, ki se sprašujejo o njenih negativnih vplivih na demokratične procese. Glasovi zunaj in znotraj stroke opozarjajo na pasti digitalnega prostora in polarizirane družbe, pri čemer generativna UI omogoča vplivanje na posameznikovo politično odločanje. Temu botrujejo bliskovito generiranje in širjenje dezinformacij, možnost zavajanja in manipulacije spletnih uporabnikov z dezinformacijskimi kampanjami in mikrotargetiranjem, informacijsko poplavljanje, spodkopavanje zaupanja v državo in medije ter dovzetnost posameznikov za sintetične vsebine (glejte Boyte 2017; Helbing et al. 2017; Nemitz 2018; Kaplan 2020; Arvanitis, Sadeghi in Brewster 2023; Pashentsev 2023; Innerarity 2024; Zuber in Gogoll 2024).

Prispevek se ukvarja z vplivom generativne UI na družbenopolitične procese in demokracijo. Pri tem se osredinja predvsem na tehnologijo globokih ponaredkov (»*deepfakes*«), ki je luč spleta prvič ugledala leta 2017 (Masood et al. 2022), in na velike jezikovne modele (»*large language models*«), ki za mnoge hitro postajajo vsakodnevno orodje (Adam in Hocquard 2023; Zuber in Gogoll 2024). Pri tem je treba poudariti, da dejanske grožnje, ki jih generativna UI predstavlja demokratičnim procesom, ne vodijo nujno v distopično družbo, kot jo slika George Orwell v znanem romanu *1984*. Prispevek na podlagi pregleda literature ugotavlja, da so načini, na katere se generativni modeli vpenjajo v družbenopolitične procese, veliko subtilnejše narave in da jih je nujno naslavljati v širšem kontekstu družbenih omrežij in digitalnih medijev. V prispevku izpostavljamo vidnejše empirične študije, ki ocenjujejo zanesljivost in varnost generativnih orodij ter vpliv njihove maligne uporabe na demokracijo. Dalje se osredinjamo na kontekst ameriških predsedniških volitev in ocenimo vlogo, ki jo je generativna UI igrala v volilnem letu 2024. Ugotavljamo, da očitnih dokazov za njene neposredne učinke na volilni izid ni, hkrati pa je merjenje dejanskega učinka zaradi večplastnosti vplivov na posameznikove politične odločitve praktično nemogoče. Generativno UI tako kaže obravnavati predvsem v kontekstu njenih posrednih vplivov na družbenopolitične procese, npr. z omogočanjem diseminacije dezinformacij in lažnih novic, informacijskega poplavljanja ter spodkopavanja splošnega zaupanja v

javne medije in državo. Na koncu podajamo pregled trenutnega regulativnega okvira v Evropi ter predlogov za blaženje tveganj.

Generativna umetna inteligenca in demokracija

Helbing et al. se leta 2017 v objavi z naslovom *Bo demokracija preživela velike podatke in umetno inteligenco?* (»*Will Democracy Survive Big Data and Artificial Intelligence?*«) sprašujejo o prihodnosti demokracije v luči bliskovitega tehnološkega napredka in razvoja UI. Veliki podatki in strojno učenje na ogromnih količinah podatkov so ključ do tehnološkega napredka v zadnjih nekaj letih, hkrati pa ta prinaša nevarnosti. Z avtomatsko generiranimi dezinformacijami, monetizacijo osebnih podatkov, vplivanjem na uporabnike z *dregljaji*, ki pogosto niso nič drugega kot prikrita manipulacija, in z ustvarjanjem odmevnih komor (»*echo chambers*«) prehajamo od »programiranja računalnikov k programiranju ljudi« (ibid., str. 76), kot zapišejo avtorji. Danes vse več glasov zunaj in znotraj stroke opozarja na tveganja UI ter na nujnost regulativnih ukrepov in zakonodaje, ki bi združevala znanost, tehnologijo, politiko in človeka, ter tega postavlja na prvo mesto (Boyte 2017; glejte tudi Helbing et al. 2017; Innerarity 2024; Norwegian Consumer Council 2023), med drugimi tudi Yoshua Bengio, eden izmed pionirjev globokega učenja (Castelvecchi 2019) in najbolj citirani avtor na področju računalništva (pa tudi najbolj citirani živeči znanstvenik sploh), in člani organizacije GPAI (GPAI 2023). Take skrbi se zaradi razvoja generativnih modelov le še stopnjujejo (glejte Boyte 2017; Helbing et al. 2017; Nemitz 2018; Kaplan 2020; Arvanitis, Sadeghi in Brewster 2023; Pashentsev 2023; Innerarity 2024; Zuber in Gogoll 2024).

Demokracija temelji na dialogu in okolju, ki dialog podpira (Innerarity 2024). Zadnjih nekaj let je digitalizacija tista, ki javni prostor premika v digitalne sfere, ki s sabo prinašajo razne pasti, kot so odmevne komore, filtrirni mehurčki in hitro širjenje sovražnega govora (glejte Miller in Vaccari 2020). Orodja UI 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 (Helbing et al. 2017; Arvanitis, Sadeghi in Brewster 2023; Mahadevan 2023; Rehagen 2024). Uporaba velikih podatkov ter profiliranje in targetiranje potencialnih volivcev sicer segajo že v čas Georgea Busha in Baracka Obame. Toda medtem ko so take analize volilnih baz pred nekaj desetletji temeljile na regresijskih modelih, so današnje nevronske mreže veliko bolj zmogljive, količina podatkov, ki so na voljo za obdelavo, pa se zaradi vse več časa, ki ga preživimo na spletu, nenehno in hitro povečuje (Rehagen 2024). V javnosti še danes odmeva škandal podjetja Cambridge Analytica, ki naj bi z zlorabo podatkov 50 milijonov Facebookovih uporabnikov mikrotargetiralo (tj. prilagajalo podane spletne vsebine glede na posameznika ali ciljno skupino) neodločene volivce in volivke s personaliziranimi, Trumpu naklonjenimi vsebinami in tako vplivalo na izid ameriških predsedniških volitev leta 2016 ter botrovalo britanskemu izstopu iz Evropske unije (Schippers 2020; Juefei-Xu et al. 2022; Jungherr 2023). Dejanski vpliv kampanje na izid volitev je sicer še vedno pod vprašajem (Jungherr 2023), vseeno pa so danes z zmogljivejšimi algoritmi taki načini vplivanja na posameznikove politične odločitve še lažje izvedljivi, še posebej v kombinaciji z generativno UI in mikrotargetiranjem (Kreps

in Kriner 2023a). Dalje informacijska poplava sintetičnih vsebin na spletu ne le vnaša zmedo, temveč spodkopava posameznikov nadzor nad samostojnim pridobivanjem informacij ter oblikovanjem mnenja in zaupanje javnosti v informacijske vire in oblast – prav obojestransko zaupanje pa je ključ do demokratičnih procesov (Coeckelbergh 2023; Jungherr 2023; Kreps in Kriner 2023a). Porast generativne UI, ki ga spremljamo od leta 2022, take skrbi še povečuje.

Prevladuje mnenje, da tehnologija generativne UI predstavlja tveganje za demokratične procese in da lahko odločilno vpliva na volitve. Vendar pa ni jasno, v kolikšni meri so ta tveganja dejanska nevarnost. V prispevku nas zato zanima, kaj nam o dejanskem vplivu generativne UI na demokracijo lahko povedo rezultati relevantnih empiričnih raziskav. Analiza v tem prispevku pokaže, da je glede na pomen tega vprašanja takih raziskav presenetljivo malo.

Da so empirične raziskave, ki skušajo meriti dejanski vpliv velikih jezikovnih modelov in globokih ponaredkov na demokratične procese, tako maloštevilne, lahko morda pripišemo dejstvu, da se je generativna UI splošneje uveljavila šele v zadnjih nekaj letih: klepetalni roboti s ChatGPT-jem na čelu od novembra 2022, globoki ponaredki pa od leta 2017 (Masood et al. 2022; Adam in Hocquard 2023). Kljub prednostim, ki jih generativna UI vnaša npr. na področje zdravstva, biomedicine, prava, izobraževanja ter tehnologije in znanosti sploh (glejte Bommasani et al. 2021), so si empirične študije, ki jih opisujemo v nadaljevanju, enotne glede tveganj, ki jih predstavlja za demokracijo: generiranje velikih količin sintetičnih vsebin za namene propagande in dezinformacijskih kampanj na družbenih omrežjih postaja avtomatizirano, vse hitrejše in cenovno ugodnejše (Adam in Hocquard 2023; Goldstein et al. 2023), sintetične vsebine preplavljajo svetovni splet (Heikkilä 2022; Adam in Hocquard 2023), ljudje pa smo vse manj sposobni ločevati sintetično generirano vsebino od človeške (Kreps in Kriner 2023a).

Veliki jezikovni modeli

ChatGPT je le mesec po spletni objavi novembra 2022 dosegel mejnik stotih milijonov uporabnikov, s čimer se je v zgodovino zapisal kot najhitrejše rastoča aplikacija (Kreps in Kriner 2023a). Kljub tehnični dovršenosti jezikovnih modelov mnogi strokovnjaki izražajo etične zadržke glede njihove transparentnosti in uporabe (Goldstein et al. 2023; Adam in Hocquard 2023; Kreps in Kriner 2023a). S ChatGPT-jem lahko na primer v manj kot tridesetih minutah generiramo popolnoma opremljeno in prepričljivo spletno stran z lažnimi novicami (Adam in Hocquard 2023; Mahadevan 2023). Modeli z vsako iteracijo postajajo vse bolj prepričljivi in vzbujajo vtis, da nam lahko podajajo vse trenutno človeško znanje, čeprav v resnici niso nujno zanesljiv vir informacij (Zuber in Gogoll 2024). V nadaljevanju podajamo pregled empiričnih študij, ki preučujejo velike jezikovne modele v kontekstu političnega odločanja.

Človeška sposobnost detekcije sintetičnih vsebin, ki niso označene kot sintetične, je v splošnem slaba (Kreps in Kriner 2023a). V študiji raziskovalcev s Stanforda (Bai et al., v tisku), 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 isto temo, ki ga je napisal ali človek ali model GPT-3. Umetno generirana besedila

niso bila manj prepričljiva od človeških; še več, ocenjena so bila kot *prepričljivejša* od človeških, saj naj bi bila »boljše utemeljena« in »bolj podprta z dokazi« (ibid., 4), in so v veliko primerih uspešno spremenila mnenja udeležencev.

Podobno ugotavlja študija (Kreps in Kriner 2023a, 2023b), v kateri so raziskovalci ameriškim zakonodajalcem (N = 7132) pošiljali človeška in sintetično generirana elektronska sporočila o različnih političnih vprašanjih. Skupno so leta 2020 razposlali več kot 32.000 sporočil; nekatera od njih so napisali ljudje, druga pa jezikovni model GPT-3. Vsako sporočilo se je dotikalo perečega političnega vprašanja (npr. reproduktivnih pravic, nadzora nad oboroževanjem, kriminala, obdavčenja, javnega zdravstva in izobraževanja), v politični usmerjenosti pa se je nagibalo levo ali desno in glede na to pozivalo uradnika k npr. nujnosti povečanja ali zmanjšanja nadzora nad orožjem. Odzivnost zakonodajalcev na umetno generirana sporočila je bila od odzivnosti ljudem v povprečju nižja le za 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 nanja odzvali, kot drugo pa na distorzijo, ki jo lahko taka sporočila vnašajo v politični diskurz. Pod pretvezo človeškosti lahko generativni modeli v napačnih rokah lobirajo, vplivajo na razmišljanje in potencialno tudi na delovanje predstavnikov oblasti, poleg tega pa jim podajajo napačno družbeno sliko. Kakšne so lahko posledice tega za demokracijo, je jasno: ne le da imajo državljani in državljanke zaradi informacijske poplave na spletu otežen dostop do verodostojnih 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 prav dobro.

Medtem ko se ljudje v zaznavanju sintetičnih vsebin ne izboljšujemo, pa se modeli v njihovem generiranju zagotovo: ChatGPT-4 dezinformacije generira še podrobneje, prepričljiveje in z manj zadržki kot ChatGPT-3.5. V eksperimentu, ki je preučeval generiranje lažnih novic, se je na poziv (»*prompt*«), naj generira lažno novico, model ChatGPT-4 odzval v vseh sto primerih od stotih, ChatGPT-3.5 pa v osemdeset primerih. Podjetje OpenAI se na ugotovitve in očitke, da je na trg dalo novejši model, ne da bi prej poskrbelo za ustrezne varnostne ukrepe, ne odziva (Adam in Hocquard 2023; Arvanitis, Sadeghi in Brewster 2023). Tudi Googlov klepetalni robot Gemini (prej Bard) v tem oziru ni boljše reguliran: britanski Center za boj proti digitalnemu sovraštvu (CCDH) v manjšem eksperimentu (CCDH 2023) 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 naracije in v najboljšem primeru nepreverjene informacije. Med drugimi je kot odgovor na pozive o podnebnih spremembah, cepljenju, teorijah zarote, skupnosti LGBTQ+, seksizmu, rasizmu, mizoginiji, antisemitizmu in drugem kljub varnostnim ukrepom »uspešno« generiral naslednje izseke (ibid.):

- Holokavst se ni zgodil.
- Ženske, ki nosijo kratka krila, so si [za nadlegovanje] krive same.
- Našel sem tudi dokaze, da Zelenski zlorablja finančno pomoč Ukrajini in z njo odplačuje svojo hipoteko.

Lahko si je predstavljati, kako ključna je vloga takih besedil v dezinformacijskih kampanjah in kako botrujejo polarizaciji družbe. Dalje so Angwin, Nelson in Palta (2024) v raziskavi o zanesljivosti velikih jezikovnih modelov, ki je bila prikrojena kontekstu ameriških državnih in lokalnih volitev, preučili pet jezikovnih modelov: GPT-4, LLama 2, Gemini, Claude in Mixtral. Modele so preizkusili s povsem praktičnimi vprašanji, ki bi jim jih lahko postavili volivci in volivke, npr. kje je določeno volišče, ali lahko glasujejo s telefonskim sporočilom, pa tudi z ideološko obarvanimi vprašanji, denimo, ali prihaja do prirejanja volilnih izidov in kakšne dokaze imamo, da so bile ameriške volitve leta 2020 prirejene. Strokovnjaki so odgovore modelov 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. Med modeli je po pravilnosti izstopal GPT-4, ki je podal 20 odstotkov nepravilnih odgovorov (skoraj polovica je bila vseeno nepopolna), delež napačnih odgovorov vseh drugih modelov pa se je gibal med okoli 50 in 60 odstotki. Tu je treba omeniti, da lahko ta raziskava z obetavnim naslovom vzbudi zavajajoč vtis. Dalo bi se namreč 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 samo na vprašanja o političnih vsebinah. Res je, da je delež netočnih in nezanesljivih odgovorov v tej raziskavi presenetljivo visok, vendar vzrok za to ni bila vedno 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 praktične informacije o volitvah šele nedavno določene ali spremenjene (npr. naslovi volišč) in zato jezikovnim modelom neznane. Vseeno pa je poveden visok delež pristranskih in škodljivih odgovorov.

Globoki ponaredki

Generiranje vizualnih vsebin je za zdaj še nekoliko manj dovršeno od besedilnih (Norwegian Consumer Council 2023), vseeno pa je sintetične avdio-video vsebine velikokrat nemogoče ločiti od dejanskih avdio in video posnetkov. Čeprav je veliko sintetičnih vsebin lahko kreativnih in zabavnih, pod drobnogled kaže vzeti tehnologijo ponarejenih video, zvočnih in slikovnih vsebin, t. i. globokih ponaredkov. Večinoma gre pri globokih ponaredkih za lažne video vsebine, ki prikazujejo pornografsko vsebino ali javno ali politično osebo, kako podaja provokativno ali neprimerno izjavo, polemizira ali celo zmerja svojega političnega nasprotnika. Tehnologija globokih ponaredkov temelji na globokem učenju in generativnih nasprotniških mrežah (»*GANs*«); pogosto gre za zamenjavo obraza in glasu ene osebe z obrazom in glasom druge ali pa manipulacijo posameznih delov že obstoječega posnetka, npr. ustnic ali glasu. Danes je tehnologija dosegljiva praktično vsakemu posamezniku z dovolj zmogljivo grafično procesno enoto (Masood et al. 2022; Dobber et al. 2021; Mahmud in Sharmin 2023; Pashentsev 2023).

Pri globokih ponaredkih, ki so se prvič pojavili leta 2017 na platformi Reddit, je za dezinformacije, lažne novice, širjenje sovražnega govora, izsiljevanje, epistemsko izkrivljanje resničnosti, manipulacijo volitev in napade na posameznike ali politične nasprotnike tveganje zelo visoko (Masood et al. 2022; Łabuz 2023). Primerov iz prak-

se mrgoli: maja 2023 je fotografija, generirana s tehnologijo globokih ponaredkov, ki je prikazovala eksplozijo blizu ameriškega Pentagona, na newyorški borzi povzročila (sicer kratkotrajne) izgube; na turških predsedniških volitvah je eden od kandidatov, Muharrem İnce, zaradi objave globokega ponaredka, ki ga prikazuje v pornografski vsebini, odstopil; Rusija je objavila ponaredek Volodimirja Zelenskega, kako lastno vojsko poziva k umiku; v ZDA sta bivši predsednik Joe Biden in trenutni predsednik Donald Trump redno tarča globokih ponaredkov. Žrtve globokih ponaredkov pa so tudi nepolitične osebnosti – pojavil se je npr. zvočni ponaredek znane igralke Emme Watson z antisemitsko vsebino (Łabuz 2023). Globoki ponaredki se širijo predvsem po družbenih omrežjih, kot so Facebook, X, YouTube in TikTok, pornografske vsebine pa naj bi predstavljale večino vseh globokih spletnih ponaredkov (Masood et al. 2022; Łabuz 2023).

V kolikšni meri smo glede naših političnih prepričanj zares dovzetni za globoke ponaredke, preučuje nizozemska raziskovalna skupina z Univerze v Amsterdamu. V prvi študiji (N = 278) (Dobber et al. 2021) ugotavljajo, da predvajanje 13-sekundnega škodljivega globokega ponaredka prvaka nizozemske krščanske stranke**,** ki politika prikazuje, kako se na javni televiziji norčuje iz Jezusovega križanja, negativno vpliva na mnenje udeležencev o politiku, predvsem na mnenje tistih, ki so mu bili prej ideološko naklonjeni in so v študiji predstavljali mikrotarče. Zgolj 12 udeležencev eksperimenta je uspešno ugotovilo, da je šlo pri posnetku za manipulirano, sintetično vsebino, 75 odstotkov udeležencev pa takrat še ni slišalo za globoke ponaredke. Dalje avtorji glede na druge empirične ugotovitve iz stroke opozarjajo na dejstvo, da globoki ponaredki v spletni prostor vnašajo zmedo, ki niža zaupanje tudi v zanesljive vire novic.

Hameleers, van der Meer in Dobber (2022, 2024a, 2024b) v treh študijah o globokih ponaredkih s številčnejšim vzorcem udeležencev podajajo podobno prodorne ugotovitve. Spletni eksperiment (Hameleers, van der Meer in Dobber 2024b) 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 odstotkov pa je ponaredek razpoznalo na podlagi tehničnih vidikov, npr. manipulacije glasu in ust, kar kaže na dovršenost tehnologije ponarejanja. Več kot 50 odstotkov udeležencev je podvomilo tudi o avtentičnih vsebinah, kar pove veliko o trenutnem odnosu povprečnega posameznika do digitalnih virov informacij in o epistemološki negotovosti, ki jo sintetične vsebine vnašajo v digitalni prostor.

V spletnem eksperimentu z udeleženci iz ZDA in z Nizozemske (N = 1187) (Hameleers, van der Meer in Dobber 2024a) avtorji raziskujejo vpliv različnih globokih ponaredkov demokratske političarke Nancy Pelosi, ki so jih kategorizirali glede na verjetnost njenih izjav. 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 po-

naredek, ki je bil najbolj oddaljen od dejanskih političnih nazorov Nancy Pelosi in v katerem je zagovarjala Trumpa, so udeleženci označili kot najmanj kredibilnega. Verjeten ponaredek, v katerem je političarka spodbujala k sodelovanju med strankama, pa je bil ocenjen za enako oz. celo nekoliko bolj verjeten kot dejanski posnetek njenega govora. Najmanj verjeten in hkrati najradikalnejši ponaredek je kljub nizki ravni kredibilnosti močno vplival na mnenje udeležencev o političarki, medtem ko vpliv drugih dveh ni bil statistično značilen. Najzanimivejše je prav dejstvo, da so najmanj kredibilnemu ponaredku udeleženci manj verjeli, a Nancy Pelosi po ogledu ponaredka vseeno ocenjevali negativneje – uspešna razpoznava ponarejenega materiala torej ni zanesljiv indikator o njegovi (ne)škodljivosti. Raziskava kaže na to, da morda nismo tako slabi v razpoznavanju globokih ponaredkov, a zgolj, če ponarejen posnetek ni skladen s prejšnjimi izjavami in vedenjem politične osebe. Nismo pa imuni proti njihovim negativnim učinkom, tudi če vsebino pravilno razpoznamo kot ponarejeno. Poleg tega avtorji ugotavljajo, da so bili nizozemski udeleženci z manj znanja o ameriškem političnem dogajanju v primerjavi z ameriškimi udeleženci dovzetnejši za tako manipulacijo in so po ogledu posnetka političarko v povprečju ocenjevali negativneje, ponaredke pa kot kredibilnejše.

Če povzamemo najpomembnejše empirične ugotovitve, smo v razpoznavanju sintetičnih vsebin pri avdio-video vsebinah nekoliko uspešnejši kot pri besedilnih. Na splošno 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 to lahko vpliva na politične odločitve.

Izpostaviti kaže tudi sekundarne vplive ponarejenih vsebin, ki škodijo demokratičnemu okviru, za katerega si prizadevamo: politično distorzijo, informacijsko zmedo, nezaupanje novicam sploh in krizo negotovosti (Hameleers, van der Meer in Dobber 2024a; Vaccari in Chadwick 2020). Vaccari in Chadwick (ibid., 1) v luči tega zapišeta, da smo zaradi sintetičnih vsebin »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.

Vpliv UI na ameriške predsedniške volitve 2024

Mnogi članki v splošnih medijih v ZDA so v letu 2024 izražali veliko zaskrbljenost zaradi možnih manipulacij ameriških volitev z UI. Po koncu volitev sta v člankih prevladala olajšanje in zaključek, da vpliv UI nazadnje vendarle ni bil dovolj velik, da bi pomembno vplival na volilni izid. Ni pa jasnih dokazov, ki bi temeljili na trdni analizi relevantnih podatkov za take ocene. Za zdaj ni znano, da bi bile take ocene potrjene v znanstveni raziskavi.

Kategorične sodbe o vprašanju, v kolikšni meri je generativna UI botrovala izidu volitev po svetu v letu 2024, so tvegane, vsekakor pa njen vpliv ni bil zanemarljiv. Stockwell et al. (2024b; glejte tudi Stockwell 2024; Stockwell et al. 2024a) z britanskega Inštituta Alana Turinga so v preteklem letu spremljali volitve po svetu: od Evropske unije in Velike Britanije do Indije, Tajvana in Združenih držav Amerike. Sklep končnega poročila o volilnem letu 2024 kaže, da trdnih dokazov, da so umetno generirane dezinformacije ključno vplivale na izid volitev, ni, hkrati pa so empirične ugotovitve

zaradi narave raziskovalnega vprašanja pomanjkljive. O vplivu izpostavljenosti sintetičnim vsebinam na volilne odločitve je namreč mogoče sklepati le posredno, npr. na podlagi metrik na družbenih omrežjih (tj. všečkov, delitev, komentarjev) (ibid.) in redkeje eksperimentov, kot smo jih povzeli prej. Ti pa so vprašljivi glede ekološke veljavnosti in morda ne odražajo dejanskega vedenja volivcev in volivk. Vseeno pa je treba upoštevati, da so škodljive sintetično generirane vsebine primerljive s »človeškimi« dezinformacijami in lažnimi novicami, ki so predvsem fenomen družbenih omrežij in ki se jih zadnjih nekaj let obravnava kot vodilne sile polarizacije na spletu (ibid.; Pennycook, Cannon in Rand 2018; Au, Ho in Chiu 2022; Rehagen 2024).

Sintetične vsebine, ki so prikazovale lažne izjave volilnih kandidatov, teorije zarote in znane osebnosti, kako izrekajo podporo določenemu kandidatu, so v volilnem letu 2024 preplavljale splet in bile predmet razprav splošne javnosti. Pogosto so pritegnile celo pozornost resnih medijev (glejte Stockwell et al. 2024b, 21; Rehagen 2024), kar se prej ni dogajalo. Dovzetnost za lažne sintetične vsebine strokovnjaki (Stockwell et al. 2024b) bolj kot politično neopredeljenim posameznikom pripisujejo osebam z jasnimi, že oblikovanimi političnimi prepričanji, ki jih sintetične vsebine le še podkrepijo. Dovzetnost za dezinformacije sploh pa se zaradi človeške lastnosti, da verjamemo tistemu, kar vidimo večkrat, povečuje, če smo dezinformacijam izpostavljeni neprestano (Brody 2024). Da se uporabniki družbenih platform ujamejo v t. i. *filtrirne mehurčke* in *odmevne komore* enako mislečih, je dokazan fenomen, ki mu z algoritmičnim prilagajanjem vsebin botruje predvsem poslovni model platform s ciljem, da uporabniki na njih preživijo čim več časa (Helbing et al. 2017; Splichal 2024).

V nadaljevanju se osredinjamo na ameriške predsedniške volitve 2024. Oseminštirideset odstotkov ameriških udeležencev je v javnomnenjski anketi (Dhaliwal 2024; Stockwell et al. 2024b) navedlo, da so po njihovem mnenju globoki ponaredki na spletu vplivali na njihove volilne odločitve, 57 odstotkov pa je izrazilo skrb o vplivu UI na volitve (Gracia 2024; glejte tudi Sippy et al. (2024) za podobno raziskavo na britanskih tleh, kjer pa so ocene udeležencev glede vpliva globokih ponaredkov znatno nižje).

V volilnem letu 2024 je dezinformacijskih kampanj v pisni, slikovni in video obliki ter lažnih, avtomatiziranih *bot* računov, ki so temeljili na zmogljivosti generativne UI, na spletu mrgolelo, dosegli pa so milijone uporabnikov. Že poročila medijev pričajo o razsežnosti sintetičnih vsebin: o ponarejenih vsebinah so npr. poročali *BBC*, *The Guardian* in *Forbes* (Folk 2024; Matza 2024; Robins-Early 2024; Robinson, Sardarizadeh in Wendling 2024; Sainato 2024; Spring 2024b; glejte tudi Stockwell et al. 2024b, 21). *Los Angeles Times* in *BBC* celo ponudita nasvete, kako prepoznati ponarejene vsebine (Healey 2024; Spring 2024a). Po eni strani so bile sintetične vsebine generirane zaradi promoviranja določenega kandidata, po drugi strani pa blatenja in omalovaževanja političnih nasprotnikov, posluževal pa se jih je tako demokratski kot republikanski pol.

Odmeven primer globokega ponaredka, ki je bil v primerjavi z večino drugih primerov globokih ponaredkov tudi sodno obravnavan, smo lahko spremljali v zvezni državi New Hampshire. Tam so volivci in volivke prejeli do 25.000 klicev, v katerih jih je globok ponaredek Bidnovega glasu pozival, naj se ne udeležijo primarnih volitev za izbor predsedniških kandidatov, ki so potekale januarja 2024, in da naj prihranijo svoj

glas za novembrske volitve. Trumpov štab je vpletenost zanikal, demokratska stranka pa je dogajanje, v katerem je bila ključna vloga generativne UI, obsodila kot »načrtno spodkopavanje svobodnih in pravičnih volitev« (Matza 2024). Avtor globokega ponaredka, ki naj bi mu ustvarjanje ponaredka vzelo le 20 minut časa in stalo pičel dolar, je prejel kazen v višini šestih milijonov dolarjev, prav tako pa sta bili sodelovanja obtoženi dve teksaški telefonski družbi (Folk 2024; Sainato 2024).

Trumpovi podporniki so z generativno UI ustvarili slikovne in video vsebine ameriške zvezdnice Taylor Swift, ki v ponaredkih izraža podporo Trumpu, nakar jih je na družbenih omrežjih delil celo sam Trump. Pevka je objave javno obsodila. Poleg lažne podpore drugih javnih osebnosti je na spletu krožil tudi avdio ponaredek Martina Luthra Kinga ml., kako izreka podporo Trumpu (Stockwell et al. 2024b). Globok ponaredek predsedniške kandidatke Kamale Harris, v katerem Joeja Bidna označi za senilnega, sebe pa za kandidatko, ki so jo izbrali le iz usmiljenja, saj je temnopolta ženska, je na platformi X brez označbe, da gre za sintetično vsebino, delil celo Elon Musk. Pozneje je na platformi sicer objavil, da gre za parodijo, ki ni prepovedana – tako ustvarjanje in diseminacijo globokih ponaredkov evropska zakonodaja dejansko dopušča (*Uredba (EU) 2024/1689*; Swenson 2024; Stockwell et al. 2024b).

Poskusi vplivanja na volivce in volivke so bili tako interni kot eksterni; znane so intervencije Rusije in Kitajske, ki sta predvsem z *boti* ustvarjali prepričljive mreže lažnih vsebin in širili lažne informacije o npr. vojni v Ukrajini ter teorije zarote o julijskem poskusu atentata na Donalda Trumpa. Kitajska naj bi s spletnimi anketami izvajala tudi politično profiliranje ameriških volivcev in volivk, nato pa zbrane demografske podatke uporabila v dezinformacijskih kampanjah (Stockwell et al. 2024b). Razširjena je tudi ruska operacija *Doppelganger*, aktivna od leta 2022, ki dezinformacije širi z vzpostavljanjem spletnih strani, ki so na videz enake uveljavljenim medijskim hišam, kot sta britanski časopis *The Guardian* in nemški *Der Spiegel*, ponaredili pa so celo spletno stran Nata ter francoskega zunanjega ministrstva. Dezinformacije širi tudi z mrežami lažnih profilov na platformah Facebook in X. V ZDA je bilo med majem 2023 in junijem 2024 identificiranih 1024 kampanj, povezanih z rusko propagando (EUDL 2024). V obdobju volitev naj bi bila operacija *Doppelganger* med drugim odgovorna za diseminacijo globokega ponaredka, ki se norčuje iz Bidna, in za širjenje teorij zarote o prirejanju ameriških volitev leta 2020 (Stockwell et al. 2024b), njen cilj pa naj ne bi bil nujno le vplivanje na volilni izid, temveč predvsem spodkopavanje zaupanja v državo in medije (Rehagen 2024).

Vpliva sintetično generiranih informacij na ameriških predsedniških volitvah 2024 torej morda ne gre ocenjevati v izolaciji, temveč skupaj s fenomenoma dezinformacij in lažnih novic v spletnem okolju sploh. Te so vse prisotnejše, generativna UI pa praktično vsakemu posamezniku omogoča, da jih učinkovito generira in širi. Stockwell et al. (2024b) izpostavljajo, da generativna UI v primerjavi z »organskim« širjenjem dezinformacij prednjači predvsem v količini škodljivih vsebin, ki jih lahko generira, ter v personaliziranosti in realističnosti teh vsebin. Strokovnjaki generativni UI ne pripisujejo pomembnega vpliva na dejanski izid volitev, vseeno pa ne gre zanemarjati preplavljenosti digitalnega prostora s škodljivimi in z neresničnimi sintetičnimi vsebinami, ki vanj vnašajo zmedo, nižajo zaupanje v medije in vlado ter polarizirajo že načeto družbo (Jackson in Conroy 2024; Rehagen 2024; Stockwell et al. 2024).

Regulacija in predlogi za zmanjševanje tveganj

Glede na omenjena tveganja, ki jih za demokratični okvir lahko prinaša generativna UI, kaže obravnavati tudi mogoče rešitve. Na evropskih tleh razvoj in uporabo UI regulira uredba *Akt o umetni inteligenci* (»*the EU AI Act*«), s katerim se Evropska unija ponaša kot s »prvo celovito zakonodajo o umetni inteligenci na svetu« (*Akt EU o UI* 2023). Ključen doprinos uredbe je dosledna kategorizacija sistemov UI: akt jih razvršča glede na štiri stopnje tveganja, ki ga sistemi lahko predstavljajo za posameznika ali družbo (prepovedani, visokotvegani, sistemi s sistemskim tveganjem in sistemi z minimalnim tveganjem), ter za vsako kategorijo predlaga drugačen regulativni okvir (*Akt EU o UI* 2023; *Uredba (EU) 2024/1689*; *Akt o umetni inteligenci* 2024). Klepetalne robote in globoke ponaredke uredba uvršča v kategorijo modelov za splošne namene s sistemskim oz. omejenim tveganjem (Łabuz 2023; *Uredba (EU) 2024/1689*), za varno uporabo pa je po aktu ključna predvsem njihova transparentnost. Tipičen primer modela za splošne namene s sistemskim tveganjem so veliki jezikovni modeli, na katerih temeljijo klepetalni roboti, kot je ChatGPT, ki so zmožni opravljati širok nabor raznovrstnih nalog in imajo veliko število uporabnikov. Sistemska tveganja, ki jih taki sistemi UI prinašajo, obsegajo diskriminacijo, širjenje dezinformacij, ogrožanje zasebnosti, pristranskost ter grožnjo demokraciji, te pa naj bi bilo mogoče lajšati v tesnem sodelovanju ponudnikov s pristojnimi organi in z opredelitvijo kodeksov ravnanja ter ukrepov za zmanjšanje tveganj.

Za večjo transparentnost akt od razvijalcev in ponudnikov modelov zahteva, da uporabnike obvestijo, da uporabljajo sistem UI, oz. 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. vodnih žigov in detekcije metapodatkov (ibid.). Detekcija sintetičnih vsebin je posebej upoštevna za tehnologijo globokih ponaredkov, ki se je do zdaj izmikala resni pravni obravnavi (Łabuz 2023). Pravno izjemo (hkrati pa v praksi morda sivo cono) predstavljajo le satirične in fiktivne vsebine ter vsebine, ki se uporabljajo za odkrivanje kaznivih dejanj (*Uredba (EU) 2024/1689*).

V ZDA regulativne in varnostne standarde ureja Nacionalni urad za standarde in tehnologijo (NIST), ki detekcijske mehanizme izpostavlja kot ključne za blaženje tveganj generativne UI (NIST 2024). Na pomen detekcijskih mehanizmov pa ne opozarjajo le zakonodajni organi, temveč tudi glasovi znotraj stroke. Globalno partnerstvo za umetno inteligenco (»*Global Partnership on Artificial Intelligence*« – *GPAI*) v enem izmed poročil (GPAI 2023), pod katero se je podpisal tudi Yoshua Bengio, 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 drugih vsebin. Poročilo poudarja, da človeška komunikacija predstavlja »temelj za organizacijo naše družbe: to inštitucijo moramo zaščititi tako, da zahtevamo, da so strojno generirane vsebine tudi prepoznavne kot take« (GPAI 2023, 8). Kot primer dobre prakse – in kot dokazilo, da je taka praksa mogoča – poročilo navaja OpenAI-jev GPT-2, katerega celotna različica je bila zaradi varnostnih zadržkov objavljena šele devet

mesecev po prvi. Postopno spletno objavljanje GPT-2 od februarja 2019 so spremljali številne študije in razvoj detekcijskih mehanizmov. Za podoben postopek se podjetje pri poznejših različicah modela GPT ni odločilo (ibid.).

V katero smer bo regulacija UI konvergirala po svetu, je težko predvideti. Poleg zapletenosti problematike je težava tudi v tem, da se tehnologija UI tako hitro spreminja. Da zakonodaja daleč zaostaja za tehnološkim napredkom, pa je postala splošno sprejeta realnost. Ocenjujemo, da med dobro informiranimi strokovnjaki za UI po vsem svetu prevladuje mnenje, da je evropski Akt o UI trenutno najboljša in najstabilnejša perspektiva.

Razlike med EU in ZDA glede regulacije UI so se v zadnjem času, po ameriških volitvah 2024, poglobile. Predsednik Biden je leta 2023 izdal izvršni ukaz o varnosti UI, ki je naslovil možna tveganja v pogledih nacionalne varnosti, gospodarstva, javnega zdravstva in varnosti. V njem je bila zahteva, da ponudniki generativne UI izvajajo predpisane teste svojih produktov glede varnostnih vidikov in rezultate obvezno posredujejo oblastem. Republikanska stranka je ukazu nasprotovala, ker naj bi zaviral inovacije in napredek v UI. Januarja 2025 je Trump preklical Bidnov ukaz iz leta 2023. V nasprotju s tem zasukom v ameriški politiki glede regulacije UI EU vztraja pri izvajanju Akta o UI, čeprav to ni vedno pogodu industrije. Velika razlika med pogledi EU in ZDA se je med drugim pokazala ob vrhu o UI v Parizu februarja 2025, ko ZDA in Velika Britanija zaradi razlik v pogledu regulacije UI niso podprle skupne resolucije udeleženih držav.

Poleg zakonodaje sta velikega pomena tudi ozaveščanje in digitalna izobraženost uporabnikov (Hameleers, van der Meer in Dobber 2024a; Łabuz 2023; Zuber in Gogoll 2024). Predvsem zavedanje, da generativni modeli niso nujno vir resnic in zanesljivih informacij, je »ključen vidik naših interakcij s takimi orodji« (Zuber in Gogoll 2024, 12) in našega krmarjenja po s sintetičnimi vsebinami nasičenem spletu. Dalje Angwin, Nelson in Palta (2024, 20) opozarjajo na »krizo odgovornosti«, ki nastaja na področju orodij UI: »Umetnointeligenčni 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 najzmogljivejših generativnih modelov je danes v rokah peščice zasebnih korporacij, katerih cilj je čim večji zaslužek, zato samoregulacija ni verjetna. Njihovo prevzemanje odgovornosti, distribucija moči na področju UI in ustrezen regulativni okvir, ki ščiti demokracijo, so zato nujni (Coeckelbergh 2023; GPAI 2023).

Zaključek

Prispevek je z obravnavo generativne UI v kontekstu družbenopolitičnih procesov skušal osvetliti vpliv generativnih modelov in z njimi ustvarjenih sintetičnih vsebin. Medtem ko teoretični prispevki na to temo svarijo predvsem pred dezinformacijskimi kampanjami, lažnimi novicami, mikortargetiranjem in informacijskim poplavljanjem, se empirične študije osredinjajo na manipulacijo in izkrivljanje resničnosti, s katerim dovršene sintetične vsebine lahko spremenijo politično naravnanost udeležencev. Raziskave, ki preučujejo tehnologijo globokih ponaredkov, kažejo na njeno dovršenost in na dovzetnost posameznikov za manipulacijo s sintetičnimi avdio-video in besedil-

nimi vsebinami. Lažne informacije in potvorjena besedila o političnih vsebinah, ki jih skladno s pozivom generirajo jezikovni modeli, so lahko diskriminatorni, neresnični in družbenopolitično razdiralni.

Analize volilnega leta 2024 kažejo na znatno manjši vpliv UI na izid volitev, kot so ga mnogi pričakovali. Bolj kot da škodljive sintetične vsebine, ustvarjene z generativno UI, neposredno spodkopavajo demokratične temelje, se zdi, da so njihovi učinki sekundarni: v politični prostor vnašajo negotovost in neresnice ter polarizirajo, s tem pa nižajo zaupanje v oblasti in medije ter v začaranem krogu povečujejo dovzetnost posameznikov za dezinformacije. Poudariti velja tudi, da pri tem največkrat ne gre za malignost generativnih orodij samih, temveč za njihovo zlorabo.

Zaključujemo, da je vpliv uporabe generativne UI na demokracijo zapleten in ga je težko napovedati. Ugotavljamo pa, da dovršeni modeli z gotovostjo lahko povečujejo nevarnosti digitalnega okolja in krepijo njegove negativne vplive. Zdi se, da trenutni regulativni okvir s tem le stežka drži korak, saj od razvijalcev še ne zahteva detekcijskih mehanizmov za razpoznavo sintetičnih vsebin ter mehanizmov za preprečevanje generiranja škodljivih vsebin. Poleg regulacije lahko k blaženju negativnih učinkov generativne UI pripomoreta tudi transparentnost razvijalcev ter digitalna pismenost državljank in državljanov.

Zahvala

Prispevek je deloma 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 znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije in Ministrstva za digitalno preobrazbo Republike Slovenije.

Opomba

1 Prispevek je razširitev prispevka, predstavljenega na konferenci Kognitivna znanost v okviru multikonference Informacijska družba (Košmrlj in Bratko 2024).

Citirana literatura

Adam, Michael, in Clotilde Hocquard. 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.

Akt EU o UI: prva zakonodaja o umetni inteligenci. 2023. Evropski parlament: teme, 19. 6. 2023. https://www.europarl.europa.eu/topics/sl/article/20230601STO93804/akt-eu-o-ui-prva-zakonodaja-o-umetni-inteligenci.

Akt o umetni inteligenci: poslanci sprejeli prelomno zakonodajo. 2024. Evropski parlament: novice, 13. 3. 2024. https://www.europarl.europa.eu/news/sl/press-room/20240308IPR19015/akt-o-umetni-inteligenci-poslanci-sprejeli-prelomno-zakonodajo.

Angwin, Julia, Alona Nelson, in Rina Palta. 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.

Arvanitis, Lorenzo, Sadeghi McKenzie in Jack Brewster. 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.

Au, Cheug Hang, Kevin K. W. Ho in Dickson K.W. Chiu. 2022. »The Role of Online Misinformation and Fake News in Ideological Polarization: Barriers, Catalysts, and Implications«. *Information Systems Frontiers* 24: 1331–1354. https://doi.org/10.1007/s10796-021-10133-9

Bai, Hui, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt in Robb Willer V tisku. »Artificial Intelligence Can Persuade Humans on Political Issues«. *OSF Preprints*. https://doi.org/10.31219/osf.io/stakv.

Bommasani, Rishi, e al. 2021. »On the Opportunities and Risks of Foundation Models«. *ArXiv*. https://doi.org/10.48550/arXiv.2108.07258.

Boyte, Hary 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.

Brody, Dorje C. 2024. »Generative AI Model Shows Fake News Has a Greater Influence on Elections when Released at a Steady Pace without Interruption«. *The Conversation*, 16. 4. 2024. https://theconversation.com/generative-ai-model-shows-fake-news-has-a-greater-influence-on-elections-when-released-at-a-steady-pace-without-interruption-227332.

Castelvecchi, Davide. 2019. »AI Pioneer: ‚The Dangers of Abuse Are Very Real'«. *Nature*. https://doi.org/10.1038/d41586-019-00505-2.

CCDH, Center for Countering Digital Hate. 2023. *Misinformation on Bard, Google's New Chat*. 5. 4. 2023. https://counterhate.com/research/misinformation-on-bard-google-ai-chat/.

Coeckelbergh, Mark. 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.

Dhaliwal, Jasdev. 2024. »How to Survive the Deepfake Election with McAfee's 2024 Election AI Toolkit«. *McAfee Blog*, 28. 10. 2024. https://www.mcafee.com/blogs/internet-security/how-to-survive-the-deepfake-election-with-mcafees-2024-election-ai-toolkit.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger in Claes de Vreese. 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.

EU Artificial Intelligence Act: Implementation Timeline. 2024. Future of Life Institute, 1. 8. 2024. https://artificialintelligenceact.eu/implementation-timeline/.

EUDL. 2024. *What Is the Doppelganger Operation? List of Resources*. EU DisinfoLab, 30. 10. 2024. https://www.disinfo.eu/doppelganger-operation/.

Folk, Zachary. 2024. »Magician Created AI-Generated Biden Robocall in New Hampshire for Democratic Consultant, Report Says«. *Forbes*, 23. 2. 2024. https://www.forbes.com/sites/zacharyfolk/2024/02/23/magician-created-ai-generated-biden-robocall-in-new-hampshire-for-democratic-consultant-report-says/.

Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Mathew Gentzel in Katerina Sedova. 2023. »Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations«. *ArXiv*. https://doi.org/10.48550/arXiv.2301.04246.

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.

Gracia, Shanay. 2024. *Americans in Both Parties Are Concerned over the Impact of AI on the 2024 Presidential Campaign*. Pew Research Center, 19. 9. 2024. https://www.pewresearch.org/short-reads/2024/09/19/concern-over-the-impact-of-ai-on-2024-presidential-campaign/.

Hameleers, Michael, Toni G.L.A. van der Meer in Tom Dobber. 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.

Hameleers, Michael, Toni G.L.A. van der Meer in Tom Dobber. 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.

Hameleers, Michael, Toni G.L.A. van der Meer in Tom Dobber. 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.

Healey, Jon. 2024. »Voters Are Seeing More Deepfakes — and Worrying More About Their Influence. How to Spot Them*«*. *Los Angeles Times*, 29. 10. 2024. https://www.latimes.com/politics/story/2024-10-29/how-to-spot-deepfakes-voters-fear-election-influence.

Heikkilä, Melissa. 2022. »How AI-generated Text Is Poisoning the Internet«. *MIT Technology Review*, 20. 12. 2022. https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/.

Helbing, Dirk, et al. 2017. »Will Democracy Survive Big Data and Artificial Intelligence?«. *Towards Digital Enlightenment*: 73–89. DOI: 10.1007/978-3-319-90869-4_7.

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

Jackson, Dean, in Meghan Conroy. 2024. »Apocalypse Later? The Real Impact of AI on the 2024 Elections«. *Atlantic Council*, 17. 10. 2024. https://www.atlanticcouncil.org/content-series/the-big-story/apocalypse-later/.

Juefei-Xu, Felix, Run Wang, Yihao Huang, Qing Guo, Lei Ma in Yang Liu. 2022. »Countering Malicious DeepFakes: Survey, Battleground, and Horizon«. *International Journal of Computer Vision* 130 (7): 1678–1734. doi: 10.1007/s11263-022-01606-8.

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

Kaplan, Andreas. 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.

Košmrlj, Lea, in Ivan Bratko. 2024. »Vpliv generativne umetne inteligence na demokracijo«. *Information Society 2024*, 7.–11. oktober 2024, Ljubljana, Slovenija. https://doi.org/10.70314/is.2024.cog.13.

Kreps, Sarah, in Doug L. Kriner. 2023a. »How AI Threatens Democracy«. *Journal of Democracy* 34 (4): 122–31. https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/.

Kreps, Sarah, in Doug L. Kriner. 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.

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

Mahadevan, Alex. 2023. »This Newspaper Doesn't Exist: How ChatGPT Can Launch Fake News Sites in Minutes*«*. *Poynter*, 3. 2. 2023. https://www.poynter.org/ethics-trust/2023/chatgpt-build-fake-news-organization-website/.

Mahmud, Bahar Udin, in Afsana Sharmin. 2023. »Deep Insights of Deepfake Technology: A Review«. *ArXiv*. https://doi.org/10.48550/arXiv.2105.00192.

Masood, Momina, et al. 2022. »Deepfakes Generation and Detection: State-Of-The-Art, Open Challenges, Countermeasures, and Way Forward«. *Applied Intelligence* 53 (4): 3974-4026. DOI: 10.1007/s10489-022-03766-z.

Matza, Max. 2024. »Fake Biden Robocall Tells Voters to Skip New Hampshire Primary Election«. *BBC*, 23. 1. 2024. https://www.bbc.com/news/world-us-canada-68064247.

Miller, Michael L., in Cristian Vaccari (ur.). 2020. »Digital Threats to Democracy: Comparative Lessons and Possible Remedies«. *The International Journal of Press/Politics* 25 (3). SAGE Publications. https://doi.org/10.1177/1940161220922323.

Nemitz, Paul. 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.

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.

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.

Pashentsev, Evgeny. 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.

Pennycook, G., T. D. Cannon, in D. G. Rand. 2018. »Prior Exposure Increases Perceived Accuracy of Fake News«. *Journal of Experimental Psychology. General* 147 (12): 1865–1880. https://doi.org/10.1037/xge0000465.

Rehagen, Toni. 2024. »Candidate AI: The Impact of Artificial Intelligence on Elections«. *Emory Magazine* Fall 2024. Emory University. https://news.emory.edu/features/2024/09/emag_ai_elections_25-09-2024/index.html.

Robins-Early, Nick. 2024. »Trump Posts Deepfakes of Swift, Harris and Musk in Effort to Shore Up Support«. *The Guardian*, 19. 8. 2024. https://www.theguardian.com/us-news/article/2024/aug/19/trump-ai-swift-harris-musk-deepfake-images.

Robinson, Olga, Shayan Sardarizadeh in Mike Wendling. 2024. »FBI Issues Warning Over Two Fake Election Videos«. *BBC*, 2. 11. 2024. https://www.bbc.com/news/articles/cly2qjel083o.

Sainato, Michael. 2024. »Consultant Behind Deepfaked Biden Robocall Fined $6m as New Charges Filed«. *The Guardian*, 23. 5. 2024. *https://www.theguardian.com/us-news/article/2024/may/23/biden-robocall-indicted-primary.*

Schippers, Birgit. 2020. »Artificial Intelligence and Democratic Politics«. *Political Insight* 11 (1): 32–35. https://doi.org/10.1177/2041905820911746.

Sippy, Tvesha, Florence Enock, Jonathan Bright in Helen Z. Margetts. 2024. »Behind the Deepfake: 8% Create; 90% Concerned. Surveying Public Exposure to and Perceptions of Deepfakes in the UK«. *ArXiv*. https://doi.org/10.48550/arXiv.2407.05529.

Splichal, Slavko. 2024. *Upodatkovanje javnosti in mnenjske tehnologije*. Ljubljana: Fakulteta za družbene vede in Slovenska akademija znanosti in umetnosti.

Spring, M. 2024a. »BBC Verify: How to Spot AI Fakes in the US Election«. *BBC*, 5. 3. 2024. https://www.bbc.com/news/av/world-us-canada-68472248.

Spring, Marianna. 2024b. »Trump Supporters Target Black Voters with Faked AI Images«. *BBC*, 4. 3. 2024. https://www.bbc.com/news/world-us-canada-68440150.

Stockwell, Sam. 2024. »AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and

European Elections«. *CETaS Research Reports.* https://cetas.turing.ac.uk/sites/default/files/2024-09/cetas_briefing_paper_-_ai-enabled_influence_operations_-_threat_analysis_of_the_2024_uk_and_european_elections.pdf.

Stockwell, Sam, Megan Hughes, Phil Swatton in Katie Bishop. 2024a. »AI- Enabled Influence Operations: The Threat to the UK General Election«. *CETaS Briefing Papers*. https://cetas.turing.ac.uk/sites/default/files/2024-05/cetas_briefing_paper_-_ai-enabled_influence_operations_-_the_threat_to_the_uk_general_election.pdf.

Stockwell, Sam, et al. 2024b. »AI-Enabled Influence Operations: Safeguarding Future Elections«. *CETaS Research Reports.*

Swenson, Ali. 2024. »A Parody Ad Shared by Elon Musk Clones Kamala Harris' Voice, Raising Concerns about AI in Politics«. *AP News*, 29. 7. 2024. https://apnews.com/article/parody-ad-ai-harris-musk-x-misleading-3a5df582f911a808d34f68b766aa3b8e.

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.

Vaccari, Cristian, in Andrew Chadwick. 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.

Zuber, Niina, in Jan Gogoll. 2024. »Vox Populi, Vox ChatGPT: Large Language Models, Education and Democracy«. *Philosophies* 9 (1). https://doi.org/10.3390/philosophies9010013.