JAHR | Vol. 15/2 | No. 30 | 2024

Izvorni znanstveni članak / Original Scientific Article
Zaprimljen / Received: 21. 10. 2024.

https://doi.org/10.21860/j.15.2.8

Ivan Cerovac[*,**], Helena Drmić[***]

Lažne vijesti, digitalne tehnologije i erozija realizacije individualne autonomije u svjetlu Kantove etike[1]

SAŽETAK

Digitalne tehnologije radikalno mijenjaju epistemičko okruženje u kojem se građani nalaze. Oblikujući kako građani prikupljaju informacije, kako komuniciraju ili kako donose odluke, digitalne tehnologije formiraju nove epistemičke prakse koje otvaraju prostor za neke stare, kao i za nove oblike manipulacije. Rad započinje analizom Kantova pojma autonomije volje te pokazuje kako i u kojim slučajevima ova autonomija može biti ugrožena. Nastavlja pružajući uvide kako digitalne tehnologije mogu ugroziti autonomiju građana te analizira sposobnost algoritama umjetne inteligencije da kroz mikrociljanje i sustave preporuka šire lažne vijesti i političku propagandu. Nadalje, rad razmatra štetan utjecaj ovih tehnologija na epistemičke prakse građana, naglašavajući tendenciju algoritama umjetne inteligencije da dovode do stvaranja epistemičkih balona ili do pretjeranog oslanjanja na velike jezične modele, pri čemu dolazi do slabljenja individualne sposobnosti prosuđivanja. Završno se razmatraju neki modeli regulacije ovih tehnologija te se ističe Kantovsko uporište za opravdanje takvih praksi.

Ključne riječi: sustavi preporuka, mikrociljanje, veliki jezični modeli, lažne vijesti, epistemički baloni, heteronomija volje.

[*] Filozofski fakultet, Sveučilište u Rijeci, Sveučilišna avenija 4, 51000 Rijeka, Hrvatska.

[**] Teološki fakultet, Sveučilište u Ljubljani, Poljanska cesta 4, 1000 Ljubljana, Slovenija. ORCID: https://orcid.org/0000-0003-3416-5295

[***] Nezavisni istraživač. Rijeka, Hrvatska. E-pošta: helena.drmic2@gmail.com

Adresa za korespondenciju: Ivan Cerovac, Filozofski fakultet, Sveučilište u Rijeci, Sveučilišna avenija 4, 51000 Rijeka, Hrvatska. E-pošta: icerovac@uniri.hr

[1] Rad je nastao u sklopu istraživačkog projekta „The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights“ koji je financiralo Sveučilište u Ljubljani, projekata „Epistemic Democracy in a Digital Era“ (IP-2024-05-4113) kojeg je financirala Hrvatska zaklada za znanost te projekata „Epistemička demokracija u digitalno doba“ (uniri-iskusni-human-23-141-3101) i „Pravo na pogrešku. Javno opravdanje javnih odluka i javnih vrijednosti“ (uniri-iskusni-human-23-154-8274) koje je financiralo Sveučilište u Rijeci.

UVOD

Digitalne tehnologije imaju snažan utjecaj na epistemičko okruženje u kojem se građani nalaze. One velikim dijelom određuju kako građani primaju nove informacije, kako međusobno komuniciraju, kako formiraju svoja vjerovanja i kako sudjeluju u političkom odlučivanju. Bilo da je riječ o praćenju vijesti o politici, kulturi ili sportu, o razgovoru s prijateljima ili kolegama na radnom mjestu, o provođenju slobodnog vremena kroz slušanje glazbe ili gledanje serija putem *streaming* servisa, ili pak o prikupljanju znanstvene literature ili istraživanju tema koje ih zanimaju, građani se danas više nego ikad ranije oslanjaju na digitalne tehnologije pogonjene algoritmima umjetne inteligencije. Sve češće i intenzivnije oslanjanje na ove tehnologije mijenja epistemičke prakse građana, a novi načini stvaranja i prenošenja informacija omogućuju napredne oblike manipulacije narušavajući tako slobodu građana i njihovu sposobnost da opravdano prosuđuju o brojnim moralno i politički relevantnim pitanjima.

Za potrebe ovog rada fokus se stavlja na digitalne tehnologije pogonjene algoritmima umjetne inteligencije i strojnog učenja koje se koriste u oblikovanju društvene interakcije putem interneta. Strojno učenje je metoda analize podataka koja omogućava obradu golemih količina informacija te uočavanje pravilnosti i obrazaca unutar podataka bez značajnijeg navođenja od strane ljudi. Ova metoda koristi se u provođenju niza digitalnih radnji, poput pretraživanja sadržaja na internetu, stvaranja preporuka na osnovi ranijih upita, profiliranja korisnika s ciljem stvaranja plana personaliziranog oglašavanja, prepoznavanja sadržaja teksta, fotografija te videosnimki i audiosnimki, kao i za stvaranje novih sadržaja.

Autonomija se u ovom radu razmatra kao sposobnost pojedinaca da kreiraju vlastita vjerovanja te donose vlastite odluke bez prisile ili manipulacijskog utjecaja od strane drugih aktera i vanjskih pritisaka, ali i bez unutarnjih ograničenja ili pristranosti, kao što su ovisnosti ili afektivna i snažna emocionalna stanja. Polazeći od Kantove moralne filozofije, rad autonomiju shvaća kao sposobnost racionalnih subjekata da stvaraju i slijede moralne zakone koje mogu sami sebi postaviti korištenjem svojih moralnih moći (racionalnost i razložnost), a bez prekomjernih ometajućih utjecaja vanjskih čimbenika kao što su želje, emocije, politički autoriteti i drugi oblici društvenih pritisaka. Rad se fokusira na utjecaj vanjskih čimbenika koji mogu ugroziti primjenu autonomije pojedinca te razmatra mogu li se digitalne tehnologije pogonjene algoritmima umjetne inteligencije smatrati štetnim utjecajem koji ugrožava primjenu individualne autonomije.

Štetan utjecaj digitalnih tehnologija pogonjenih algoritmima umjetne inteligencije na sposobnosti građana da formiraju opravdana vjerovanja već je neko vrijeme u fokusu rasprava unutar političke filozofije (Allcott i Gentzkow, 2017; Coeckelbergh,

2023; Hao, 2019). Budući da suvremena liberalna i demokratska društva svim građanima daju (barem u formalnoj političkoj sferi) jednaku mogućnost utjecanja na konačne ishode političkih procesa, epistemička kvaliteta demokratskih procedura u značajnoj će mjeri ovisiti i o sposobnostima građana koji u tim procedurama odlučivanja sudjeluju (Cerovac, 2022; Mill, 1977). Međutim, budući da sposobnosti građana i njihova mogućnost da ih ispravno koriste ovise o epistemičkom okruženju unutar kojeg djeluju, a to je okruženje podložno utjecaju ekonomske i društvene moći (primjerice, kroz financiranje političkih kampanja, plaćene oglase u medijima, financiranje *think tankova* koji istražuju i zagovaraju neke političke ideje), kvalitetu demokratskih procedura moguće je ugroziti kroz manipuliranje građanskim promišljanjem, formuliranjem političkih stavova i preferencija te samim glasanjem (Cerovac, 2023; Christiano, 2010). Politički filozofi se tako bave proučavanjem utjecaja digitalnih tehnologija na prosuđivanje pojedinaca prvenstveno kroz mogućnosti manipulacije koje te tehnologije pružaju te kroz analizu štetnih utjecaja manipulacije na kvalitetu procedura donošenja odluka.

Djelovanje digitalnih tehnologija na spoznajne procese građana također je već neko vrijeme u fokusu socijalne epistemologije. Promjene koje je prošireno korištenje ovih tehnologija uvelo u načine kako se građani informiraju i međusobno komuniciraju potiču razvoj štetnih epistemičkih pojava i okruženja kao što su komore jeke i epistemički baloni (Kiri Gunn, 2021; Nguyen; 2020; Sunstein, 2009) te facilitiraju širenje dezinformacija i lažnih vijesti (Consentino 2020, Rhodes 2022). Dosadašnja rasprava, međutim, u značajnoj mjeri zanemaruje analizirati kako digitalne tehnologije utječu na autonomiju pojedinaca i na njihovu sposobnost da donose odluke na temelju točnih i provjerenih informacija te ne budu žrtve manipulacije[2].

Socijalni se epistemolozi tako usmjeruju na štetan utjecaj manipulacije u procesu kolektivne potrage za istinitim (ili barem opravdanim) vjerovanjima, dok se politički filozofi fokusiraju na štetan utjecaj manipulacije na kvalitetu demokratskih odluka. U oba se slučaja zanemaruju moralni aspekti manipulacije i njezino potkopavanje digniteta drugih osoba. Uz to, budući da ne postoji uspostavljeni evaluativni okvir za procjenjivanje učinaka koje digitalne tehnologije imaju na autonomiju pojedinaca, nedostaju i obuhvatni prijedlozi zakonske regulative te javnih politika i mjera kojima bi se uklonile potencijalne štetne posljedice koje digitalne tehnologije pogonjene umjetnom inteligencijom imaju na prosuđivanje pojedinaca.

[2] Iako je ova rasprava relativno zapostavljena, postoji nekoliko primjera radova koji zahvaćaju odnos digitalnih tehnologija i autonomije pojedinaca. Primjerice, Moles (2007) u svojoj doktorskoj disertaciji analizira mentalnu kontaminaciju i proučava kako dezinformacije oblikuju stavove i odluke pojedinaca koji su im izloženi. Sahebi i Formosa (2022) fokusiraju se isključivo na društvene mreže i proučavaju njihov ograničavajući utjecaj na autonomiju pojedinaca, dok Cerovac i Drmić (2023) raspravljaju o lažnim vijestima i njihovom štetnom utjecaju na građansku sposobnost praktičnog rasuđivanja.

Glavna je teza ovog rada da digitalne tehnologije ugrožavaju primjenu individualne autonomije kroz potkopavanje epistemičkih sposobnosti građana te na taj način sprečavaju ili otežavaju moralno odlučivanje, budući da je sloboda preduvjet za moralnu odgovornost. Njegov inovativni doprinos obuhvaća promjenu fokusa dosadašnje filozofske rasprave o utjecaju umjetne inteligencije. Umjesto naglašavanja pitanja vezanih uz demokraciju ili kolektivno traženje za istinom, rad propituje mogu li građani uopće djelovati moralno ako žive i djeluju u narušenom epistemičkom okruženju.

Rad je podijeljen u tri dijela. U prvom se ukratko razmatra Kantovo shvaćanje autonomije te se proučava na koji način manipulacija predstavlja napad na negativnu slobodu građana, na primjenu individualne autonomije te kršenje digniteta osobe kojom se manipulira. U drugom se dijelu obrazlaže kako digitalne tehnologije pogonjene algoritmima umjetne inteligencije mogu imati manipulativni učinak te ozbiljno ugroziti primjenu autonomije pojedinaca. Štoviše, pokazat će se da ove tehnologije mogu izmijeniti i unakaziti epistemičko okruženje u kojem se građani nalaze do te razine da oni više nisu u mogućnosti razlikovati istine od laži u digitalnom okruženju, što može dovesti do njihove pasivizacije i odustajanja od pokušaja da se spozna i razumije svijet koji ih okružuje. U trećem se dijelu analiziraju neki oblici regulacije digitalnih tehnologija te se razmatra opravdanost i učinkovitost njihove primjene.

AUTONOMIJA U KANTOVOJ MORALNOJ FILOZOFIJI

Kant na autonomiju pojedinaca ne gleda kao na njihovu sposobnost da budu samodostatni i da žive neovisno o drugima. Isto tako, o autonomiji govori općenitije i apstraktnije nego što se taj pojam upotrebljava u suvremenim bioetičkim raspravama[3] o pitanjima pobačaja, eutanazije ili o specifičnim pravima građana u liberalnim društvima[4]. Autonomija volje za Kanta predstavlja nužan preduvjet rasprave o moralu – moralne obaveze i dužnosti nas obvezuju samo ukoliko smo racionalni subjekti čija volja može biti autonomna (Kant, 2002, str. 440-443). Sama sloboda volje može se razumjeti kao sloboda shvaćena na negativni i na pozitivni način. Negativna sloboda označava odsustvo prisile ili vanjskih utjecaja koji oblikuju proces rasuđivanja kod pojedinca. Ukoliko osoba promišlja i odlučuje u strahu od kazne, društvenih normi ili političkog autoriteta, vodeći se svojim emocijama ili željama, ili u epistemičkim uvjetima u kojima se njezinim vjerovanjima o svijetu sustavno manipulira, utoliko osoba neće imati slobodu volje u negativnom smislu (Kant, 2002, str. 445). Pozitivna

[3] Kao primjer vidi Baccarini i Prijić-Samaržija (2007).

[4] Detaljniju raspravu o tome što Kant ne smatra autonomijom volje vidi u Hill (1989).

sloboda označava slobodu osobe da sama sebi postavlja zakone, da bude zakonodavac u kraljevstvu ciljeva. Osoba slobodno djeluje prema moralnim zakonima koje si je sama postavila, odnosno prema univerzalnim moralnim zakonima do kojih je sama došla upotrebom svog uma (Kant, 2002, str. 440, vidi još Reath, 2006 i Hill, 2013).

Iako se suvremene rasprave unutar moralne i političke filozofije puno više bave slobodom shvaćenom u pozitivnom smislu i povezuju je s autonomijom, ovaj se rad usmjerava na negativno shvaćenu slobodu te na utjecaj koji vanjski čimbenici mogu imati na moralno djelovanje pojedinca. Naime, uz shvaćanje autonomije kao primarno pozitivne slobode, Kant jasno upućuje da ne smijemo zanemariti negativnu slobodu. Kada si postavljamo moralno i praktično pitanje „Što trebam činiti?“, moramo imati koncepciju nas samih kao bića slobodnih u negativnom smislu, čije djelovanje nije ograničeno emocijama, autoritetima ili drugim vanjskim čimbenicima (Piper, 2024). Bez negativne slobode ne možemo se smatrati autonomnim moralnim subjektima (Kant, 2002, str. 446-447, vidi i Hill, 2013, str. 17). Ovakvo Kantovo shvaćanje autonomije (utemeljeno na interpretaciji koju daje Thomas Hill[5]) daje prostora za tumačenje autonomije ne samo kao interne sposobnosti ili potencijala, već i kao preduvjeta za primjenu moralnih zakona. To tumačenje, koje ostaje u Kantovu duhu, usmjerava pažnju i prema uvjetima primjene i ostvarenja autonomije. Fokus se tako stavlja na heteronomne čimbenike koji ugrožavaju negativnu slobodu, a koji u vrijeme suvremenih digitalnih tehnologija postaju sve izraženiji, utjecajniji i teže uočljivi.

Realizacija autonomije osobe može biti ugrožena ili otežana na više načina. Ako osoba djeluje u skladu s praktičnim umom, ali je na samu radnju motivira neka želja ili sklonost, osoba ne djeluje autonomno (iako ima sposobnost ili potencijal da svojoj volji propiše zakone koji mogu važiti za svako drugo umno biće), već heteronomno, budući da ne djeluje iz poštovanja prema moralnom zakonu, već samo u skladu s njim. Primjerice, ako osoba pomaže drugima jer se čineći to osjeća dobro, njezina će volja biti heteronomna iako se sama radnja može činiti moralnom i pohvalnom. Isto vrijedi i za slučajeve u kojima osoba djeluje slijedeći vanjski autoritet – čak i ako na ovaj način djeluje u skladu s moralnim zakonima, osoba ne prakticira autonomiju, budući da je ne motivira moralni zakon do kojeg je došla upotrebom praktičnog uma. Djelovanje iz navike motivirano prihvaćenim društvenim normama, čak i kada je u skladu s moralnim dužnostima, također neće zadovoljiti uvjet autonomnog djelovanja, budući da nije motivirano racionalnim promišljanjem o moralnim

[5] „Volja osobe s autonomijom volje je slobodna u negativnom smislu. To jest, riječ je o vrsti kauzalnosti koja može biti aktivna, neovisno o vanjskim uzrocima koji je određuju. Drugim riječima, zamišljamo osobu s negativno slobodnom voljom kao osobu sposobnu djelovati i uzrokovati događaje bez da su njezini izbori uzročno određeni prethodnim fizičkim ili psihološkim silama“ (Hill, 2013, str. 18).

zakonima, a isto neće postići ni djelovanje iz prudencijalnog rasuđivanja (u formi hipotetičkog imperativa), kada osoba djeluje u skladu s moralnim dužnostima, ali s instrumentalnom nakanom ostvarivanja nekog osobnog cilja. Završno, Kant (2002, str. 445) ističe da ni osoba čijim se promišljanjem manipulira kroz laži i obmane neće moći realizirati autonomiju volje. Ako osoba motivirana moralnim dužnostima donira novac terorističkoj organizaciji koja je uspije zavarati i lažno se prikazati kao dobrotvorna humanitarna organizacija, osoba će naizgled djelovati iz dužnosti, no zapravo neće biti slobodna u negativnom smislu. Naime, Kant eksplicitno naglašava kako, da bi autonomno prosuđivala, osoba ne smije biti žrtva obmane ili manipulacije (Kant, 2002, vidi i O'Neill, 2014).

Primjena (sposobnosti) moralne autonomije utemeljena je u sposobnosti osobe da djeluje iz moralnih zakona koje si sama može postaviti. Međutim, osoba koja je obmanuta lažnim vijestima i drugim oblicima dezinformacija, kao i osoba koja se nalazi u epistemičkom okruženju u kojem ne može razlikovati istine od laži, više ne može djelovati prema praktičnom umu, budući da djeluje prema razlozima koji nisu njezini vlastiti, već su joj nametnuti izvana. Odluke koje osoba donosi prestaju biti rezultat njezine racionalne volje i postaju rezultat vanjskih utjecaja. Kognitivna neovisnost zbog toga predstavlja bitan preduvjet realizacije autonomije volje (Zinkin, 2024).

Manipulacija koja ugrožava primjenu autonomije volje može se manifestirati na više načina. Izlaganje osobe lažima u koje će ona povjerovati jedan je od očitih primjera. Međutim, manipulacija ne mora nužno uključivati laganje – namjerno izazivanje snažnih emotivnih i afektivnih stanja u kojima osoba neće moći primjereno rasuđivati, kao i pružanje istinitih statističkih podataka s ciljem iskorištavanja pristranosti i predrasuda koje osoba ima, mogu biti učinkoviti oblici manipulacije (Coeckelbergh, 2022; Frierson, 2005). U drugom dijelu rada pokazat će se kako su digitalne tehnologije pogonjene algoritmima umjetne inteligencije izuzetno učinkovite u facilitiranju svih navedenih oblika manipulacije. Ipak, prije toga je potrebno zahvatiti zašto je manipulacija unutar okvira Kantove moralne filozofije toliko problematična.

Manipulativan utjecaj na rasuđivanje neke osobe, smatra Kant, vrijeđa istovremeno i negativnu slobodu i dostojanstvo te osobe. Manipulirajući rasuđivanjem neke osobe, čak i ako to činimo s dobrim namjerama i s ciljem da tu osobu zaštitimo, tretiramo je poput djeteta, a ne poput odrasle osobe te joj ne pridajemo poštovanje koje zaslužuje[6] (Quong, 2010). Manipulacija tako, poput laganja ili vršenja prisile,

[6] Zanimljiva rasprava vezana uz ove oblike manipulacije razvila se među teorijama javnog opravdanja unutar političke filozofije. Rawlsova koncepcija javnog opravdanja (koja se oslanja i nadahnuće crpi iz Kantove moralne filozofije) tako polazi od ideje kako građani ne smiju kao javne razloge nuditi one u koje sami ne vjeruju, ali s kojima se drugi slažu. Takvo se sudjelovanje u javnoj raspravi smatra manipulativnim i neprihvatljivim, budući da se druge građane nastoji uvjeriti u prihvaćanje neke političke odluke razlozima u koje sami ne vjerujemo. Ovakav

oduzima autonomiju žrtvi (Korsgaard, 2007; Sunstein, 2022) te je „čini podložnom volji drugih" (Margalit, 2016, str. 104). Budući da dostojanstvo proizlazi iz ljudske racionalne prirode (Kant, 2002, 2015; vidi i Eterović, 2017), a manipulirajući drugima negiramo upravo tu vrijednu (racionalnu) ljudskost u njima, manipulativnim ponašanjem ih prestajemo tretirati kao autonomne moralne subjekte i svodimo ih na puke objekte, sredstva koja služe za postizanje naših ciljeva. Čak i kad su naše namjerne dobre a ciljevi plemeniti, ovo nezaobilazno negira moralni status drugih i potkopava njihovo dostojanstvo kao ljudskih bića.

Iako nam Kantova moralna filozofija nije nužna za opisivanje i objašnjenje etičkih problema vezanih uz manipulaciju, budući da je to moguće postići i kroz druge pristupe, poput onih utemeljenih na blagostanju (Mill, 2020; vidi i Baron, 2016), ona nam omogućuje razumijevanje nekih aspekata moralne pogrešnosti manipulacije koje drugi pristupi ne pokrivaju. Uz to, pokazuje da manipulacija može imati i dalekosežne implikacije na naše moralno okruženje, budući da moralnu odgovornost mogu snositi samo slobodni moralni subjekti (Kant, 2002; vidi i Hill, 2013), pa se tako u okolnostima raširene i sustavne manipulacije facilitirane digitalnim tehnologijama pogonjenim algoritmima umjetne inteligencije dovodi u pitanje održivost moralnog sustava u suvremenim tehnološkim društvima. Ipak, da bi se argumentirano izložila opasnost koju negativnoj slobodi pojedinaca i primjeni individualne autonomije predstavljaju algoritmi umjetne inteligencije, potrebno je sagledati na koje načine oni mogu manipulirati ljudskim prosuđivanjem i razmotriti razlikuje li se ta manipulacija u značajnoj mjeri od tradicionalnih oblika propagande koji se koriste stoljećima.

DIGITALNE TEHNOLOGIJE KAO PRIJETNJA INDIVIDUALNOJ AUTONOMIJI

Epistemičko okruženje u kojem se osoba nalazi ima značajan utjecaj na njezino prosuđivanje i može facilitirati heteronomiju (umjesto autonomije) volje. Naime, iako moralna načela proizlaze iz praktičnog uma i ne ovise o specifičnim informacijama o svijetu, njihova primjena će uvelike ovisiti o informacijama kojima pojedinac raspolaže. Uz to, čak i promišljanje o moralnim načelima može biti otežano ako se osoba nalazi u dugotrajnom i sustavno manipulativnom epistemičkom okruženju. Primjerice, kako bi se među građanima opravdao neljudski tretman uspostavljen brojnim antisemitskim politikama, nacisti su 1930-ih i 1940-ih trebali razviti i proširiti niz pseudoznanstvenih teorija koje su imale za cilj dokazati (na empirijski

način uvjeravanja tako počinje sličiti ponašanju roditelja koji pokušava nagovoriti dijete da ranije otiđe spavati uvjeravajući ga kako Djed Božićnjak nagrađuje poslušnu djecu (Rawls, 2005, vidi i Quong, 2010).

način, od mjerenja lubanja i rasne biologije preko genetike i ideja o rasnoj čistoći do pseudoznanstvenih teorija o židovskoj psihologiji) da Židovi nisu ljudi na isti način na koji su to Nijemci (Ferretti, 2018). Ove deskriptivne teorije koje se bave empirijskim podacima su, u kombinaciji s propagandom i rasnom ideologijom utkanom u razne grane povijesti, antropologije, filozofije, umjetnosti i obrazovanja, svakako ugrozile primjenu autonomiju volje građana i otežale im ili onemogućile da ispravno primjenjuju moralna načela. Iako su društveni i politički uvjeti u suvremenim zapadnim demokracijama značajno drukčiji, razvoj digitalnih tehnologija i primjena algoritama umjetne inteligencije u stvaranju i diseminaciji informacija može na sličan način dovesti do heteronomije volje građana. Ovaj se dio rada bavi analizom digitalnih tehnologija i algoritama koji značajno utječu na epistemičko okruženje u kojem se građani nalaze (ili će se nalaziti) u 21. stoljeću te razmatra četiri oblika utjecaja.

(i) Lažne vijesti i dezinformacije u digitalnom prostoru

Lažne vijesti predstavljaju netočne ili obmanjujuće informacije koje su namjerno kreirane da nalikuju na tradicionalne medijske izvještaje (Allcott i Gentzkow, 2017; McIntyre, 2018). Da bi informacija bila označena kao lažna vijest, mora zadovoljiti četiri kriterija. Prvo, informacija mora biti netočna ili pouzdano dovesti slušatelja do netočnog zaključka (Gelfert, 2021; Rini, 2017). Drugo, ona mora biti stvorena s namjerom obmanjivanja. Treće, lažne vijesti moraju oponašati format tradicionalnog izvještaja kako bi stekle vjerodostojnost. Četvrto, informacije svojim sadržajem i formatom moraju imati potencijal za široku distribuciju, obično putem društvenih mreža, što se postiže senzacionalizmom i emocionalnim nabojem (Zimdars i McLeod, 2020). Brojna istraživanja potvrđuju opasnosti koje lažne vijesti mogu imati na osobne živote građana (Cerovac i Drmić, 2023; Rapp, 2016), kao i na kvalitetu demokratskih procesa (McKay i Tenove, 2021).

Nove tehnologije pogonjene algoritmima umjetne inteligencije pospješile su doseg i brzinu širenja dezinformacija u digitalnom prostoru. Društvene mreže omogućavaju milijunima korisnika da dijele informacije i vijesti, kao i dezinformacije i lažne vijesti, koje se zbog senzacionalističkog formata i emocionalnog naboja šire izrazito brzo i oblikuju rasuđivanje velikog broja građana kojima su društvene mreže glavni izvor informacija (Kiri Gunn, 2021). Uz to, ako neka osoba pokaže interes ili uđe u interakciju s lažnim vijestima, sustavi koji korisnicima preporučuju sadržaj na osnovi njihovih ranijih upita težit će nastaviti voditi osobu do novih, tematski povezanih lažnih vijesti koje će potvrditi predrasude ili pogrešna nagađanja od kojih je osoba izvorno krenula (Zhang i sur., 2021). Poznat primjer ovakve epistemičke ovisnosti o obmanjujućim izvorima informacija vezuje se uz teoriju zavjere poznatu

kao Pizzagate, u kojoj je oko 10 % registriranih glasača u SAD-u vjerovalo kako je predsjednička kandidatkinja Hillary Clinton uključena u pedofilski lanac koji sjedište ima u manjem restoranu u Washingtonu (Zimdars i McLeod, 2020). Ipak, utjecaj na brzinu i doseg širenja lažnih vijesti samo je jedan od zabrinjavajućih učinaka digitalnih tehnologija.

S razvojem naprednijih algoritama i velikih jezičnih modela, ove tehnologije sada mogu facilitirati i kreiranje lažnih vijesti, od izrade obmanjujućih mrežnih stranica u svega nekoliko sekundi (ili tisuća takvih stranica u danu) do manipuliranja internetskim tražilicama kako bi novokreirani sadržaj dobio prioritet u prikazivanju korisnicima (Endert, 2024; Spitale i sur., 2023). Mogućnost manipuliranja fotografijama i videozapisima dosegla je do sada neviđene razine, tako da i korisnici s malo tehničkog znanja mogu kreirati izrazito uvjerljive materijale koje gotovo svi građani (osim onih koji nisu posebno trenirani za to) ne mogu prepoznati kao fabricirani sadržaj. Poznat je slučaj izmijenjene i manipulirajuće snimke koja prikazuje Nancy Pelosi, članicu Zastupničkog doma Kongresa SAD-a, kako pijano tetura i govori pred medijima (Hameleers i sur., 2024), a koju je svojevremeno putem društvenih mreža (iako je bilo riječ o izmijenjenom i manipulirajućem videu) podijelio tadašnji američki predsjednik. Ovakav je sadržaj uz pomoć umjetne inteligencije danas veoma lako kreirati i proširiti društvenim mrežama, lažnim mrežnim stranicama koje imitiraju format tradicionalnih medija, kao i putem poruka u velikim grupama i kanalima platformi za komunikaciju (npr. Telegram) (La Morgia i sur., 2021). Uz to, napredak u razvoju umjetne inteligencije omogućuje stvaranje sve razvijenijih botova, automatiziranih programa koji mogu obavljati brojne zadaće na internetu, koji lažno nastupaju kao drugi korisnici (ljudi) te šire dezinformacije, utječu na javno mnijenje, manipuliraju algoritmima u društvenim mrežama te stvaraju iluziju potpore nekom kandidatu ili privatnom poduzeću, nekom proizvodu ili javnoj politici (Ferrara, 2020).

Štetan utjecaj lažnih vijesti na prosuđivanje pojedinaca prisutan je čak i kada je sama informacija ispravno i na vrijeme prepoznata kao neistinita. Naime, čak i ako građani ne vjeruju dezinformacijama koje susreću preko medija, društvenih mreža i općenito u javnoj sferi, sama činjenica da su se s njima susretali (a zahvaljujući algoritmima umjetne inteligencije, to susretanje može biti učestalo i sustavno) naštetit će njihovu rasuđivanju (Ecker i sur., 2022; Menczer i Hills, 2020). Primjerice, pojedinci koji su se susretali s lažnim vijestima oko opasnosti cijepljenja i koji su ispravno prepoznali te dezinformacije kao lažne i dalje će biti manje skloni cijepljenju nego skupina koja nije bila izložena lažnim vijestima. Autonomija pojedinaca je tako dovedena u pitanje, budući da postoji izvanjski manipulativni učinak koji usmjerava njihovo rasuđivanje i njihovu primjenu moralnih načela.

Utjecaj digitalnih dezinformacija na autonomiju pojedinaca razoran je, budući da potkopava njihovu sposobnost donošenja informiranih odluka i otvara vrata novim oblicima manipulacije. Građani se danas sve više oslanjaju na digitalne tehnologije za informiranje o događajima u svijetu, osobito na one koje nisu moderirane i nemaju strogu (ili ikakvu) uređivačku politiku. Primjerice, istraživanja provedena 2024. godine u SAD-u pokazuju da čak 54 % građana koristi društvene mreže za informiranje o događanjima u državi i svijetu, a 18 % ističe kako su im društvene mreže preferirani izvor vijesti (PEW Research Center, 2024). Opravdano je očekivati kako će taj trend nastaviti rasti i u budućnosti, kao što će rasti i sposobnosti algoritama umjetne inteligencije da kreiraju obmanjujuće i manipulirajuće sadržaje čiju pravu prirodu korisnici neće moći prepoznati. U takvom će epistemičkom okruženju građani ili vjerovati u neke od izvora vijesti, izlažući se tako u većoj ili manjoj mjeri mogućnostima manipulacije, ili će početi osjećati „epistemički sram" (Coeckelbergh, 2024, str. 1343) te se moralno i politički pasivizirati, budući da će biti svjesni kako više ne mogu razlikovati istinu od laži. Naposlijetku, kako prosuđivati i primjenjivati moralne zakone ako ne znamo što je stvarnost a što je obmana? Ova neizbježna sumnja u vlastite epistemičke sposobnosti, suprotna Kantovom (2010) pozivu „Sapere aude!" (često prevođenom kao „Usudi se koristiti vlastiti um!"), predstavlja negaciju vlastite autonomije i slobode mišljenja. Naime, sposobnosti digitalnih tehnologija da stvaraju, oblikuju i šire lažne informacije, uključujući i sposobnosti da kroz botove „glume" druge ljude u digitalnom svijetu, mogu toliko kontaminirati epistemički prostor u kojem se krećemo da epistemičke sposobnosti građana više neće biti adekvatne za odgovorno snalaženje u ovakvim okolnostima.

(ii) Epistemički baloni

Suvremena politička i socijalna epistemologija sve veću pozornost usmjeravaju na epistemičke mreže kroz koje se informacije šire unutar zajednice. Struktura ovih mreža utječe na vrstu informacija koje se šire, na brzinu širenja, na gubitke u preciznosti i točnosti sadržaja koji se događaju kada se informacija širi, kao i na utjecaj koji sama informacija ima na prosuđivanje pojedinca koji je dio neke epistemičke mreže (Singer i sur., 2021). Epistemički baloni i komore jeke dva su ključna koncepta za razumijevanje ove rasprave. Epistemički baloni predstavljaju zatvorena epistemička okruženja u kojima su pojedinci izloženi relativno uskom rasponu gledišta, najčešće zato što se različite perspektive ignoriraju ili isključuju. Nove informacije teško ulaze u balon, a neke informacije iz njega teško izlaze. Epistemički baloni često vode do političke polarizacije i drugih štetnih pojava, no njihov nastanak uglavnom nije rezultat svjesnog i namjernog djelovanja pojedinaca, već do njega dolazi kroz selektivno izlaganje informacijama. Komore jeke na sličan način ograničavaju pristup novih informacija i perspektiva, no za razliku od epistemičkih balona one nastaju

svjesno i s namjerom, primjerice kroz dogovorenu cenzuru unutar neke grupe (ili epistemičke mreže) kojom se uklanjaju ili destimuliraju različita razmišljanja ili kritičko propitivanje postojećeg konsenzusa (Kiri Gunn, 2021; Nguyen, 2020). U oba slučaja rezultat je nepovoljno epistemičko okruženje koje negativno utječe na prosuđivanje pojedinaca koji čine dio takve epistemičke mreže. Ono pojedincima otežava pristup novim informacijama, dovodi do potvrđivanja postojećih pristranosti i predrasuda, potiče političku polarizaciju te smanjuje sposobnosti kritičkog promišljanja, što građane čini još podložnijima manipulaciji, budući da ne dolaze do novih izvora informacija (Pollock, 2024).

Komore jeke nisu u fokusu ovog rada, budući da se one, iako epistemički štetne, kreiraju svjesno i namjerno te pojedinci u njih ulaze (ili ih kreiraju) svjesni cenzure i znajući kakva pravila rasprave odlikuju takve epistemičke mreže. Značajno su nam zanimljiviji epistemički baloni, budući da se u njihovu slučaju selektivno izlaganje informacijama događa nenamjerno i nesvjesno, bez puno znanja pojedinca o ograničavajućim učincima balona u kojem se nalazi (Pariser, 2011). Budući da se danas u zapadnom svijetu većina građana informira putem interneta, bilo da je riječ o čitanju članaka na portalima, praćenju društvenih mreža ili YouTube kanala i podcasta (PEW Research Center, 2024), značajan faktor u selektivnom izlaganju informacijama predstavljaju algoritmi umjetne inteligencije koji, kroz mehanizme preporuka, određuju koje će informacije, kada i u kojoj mjeri biti ponuđene krajnjem korisniku. Riječ je o individualiziranom pristupu (koji je moguć baš zbog algoritama umjetne inteligencije koji obrađuju goleme količine podataka, uključujući i podatke o korisniku i njegovoj povijesti na internetu), a koji služi kako bi zadržao ili povećao interes korisnika za ponuđene sadržaje te pojačao interakciju koju korisnik ima s njima.

Ključni problem kod algoritamski kreiranih epistemičkih mjehura je da oni predstavljaju epistemičko okruženje koje može bitno utjecati na prosuđivanje pojedinca, a samo nije pod kontrolom tog istog pojedinca (Coeckelbergh, 2023). Koristeći tražilice kao što su Google, Bing ili Yahoo, gledajući videomaterijale ili pretežući podcaste putem YouTube kanala ili provodeći vrijeme na društvenim mrežama mi se neizbježno krećemo unutar određenog epistemičkog okruženja (u ovom slučaju, okruženja koje ima format epistemičkog mjehura) koje nismo sami kreirali i na koje ne možemo utjecati, a koje pak značajno utječe na naše prosuđivanje i formiranje vjerovanja. Iako se danas, barem u liberalnim demokracijama[7], ovi algoritmi koriste prvenstveno za maksimizaciju profita tvrtki koje su ih dizajnirale, oni mogu kroz plaćeno oglašavanje lako biti iskorišteni kako bi utjecali na promišljanje pojedinaca

[7] Direktnija upotreba algoritama umjetne inteligencije koji upravljaju rezultatima internetskog pretraživanja može se naći u brojnim neliberalnim državama svijeta. Primjerice, Kina je poznata po manipuliranju tražilicama kako bi širila političku propagandu i kontrolirala epistemičko okruženje svojih građana (Zhang, 2020).

i postigli neki politički, društveni ili ekonomski cilj za koji se zalažu oglašivači (Tufekci, 2014). Na ovaj su način građani izloženi kontinuiranoj manipulaciji, kreću se kroz epistemičko okruženje koje ne mogu kontrolirati te je njihova autonomija volje ozbiljno narušena.

(iii) Mikrociljanje i sustavi preporuka

Algoritmi umjetne inteligencije koji stoje iza mehanizama preporuka te tako omogućuju stvaranje epistemičkih balona facilitiraju još jedan oblik problematične diseminacije informacija. Riječ je o korištenju algoritama za analizu i kreiranje pomno osmišljenih i prilagođenih vijesti i oglasa usmjerenih prema manjim grupama ljudi s obzirom na njihove demografske osobine, političke ili tržišne preferencije i interese te ponašanja na internetu. Ove metode (poznate kao mikrociljanje) značajno su uspješnije od standardnih oglasa distribuiranih kroz tradicionalne medije te omogućavaju iskorištavanje raznih pristranosti kod korisnika kako bi im poruke koje se putem oglasa šalju zvučale što uvjerljivije (Barbu, 2014). Mikrociljanje se tipično koristi kako bi oglasi za različite proizvode i usluge došli upravo do onih korisnika za koje se procjenjuje da će biti najviše zainteresirani za ponuđene stvari. Analizirajući velike količine podataka, uključujući osobne podatke o korisnicima i njihovo ponašanje na internetu, moguće je precizno usmjeriti plaćeni oglas ili pak prilagoditi rezultate tražilica.

Tehnološki napredak te prikupljanje i akumuliranje sve veće količine podataka o korisnicima omogućava nove i još sofisticiranije modele mikrociljanja koji se sve više koriste u političke svrhe. Djelujući kao „manipulacijski stroj" (Simchon i sur., 2024) algoritmi mikrociljanja omogućuju povezivanje različitih informacija o korisnicima kako bi stvorili što učinkovitije oglase, prilagođavajući ne samo sadržaj oglasa već i njegov jezik i stil izražavanja, vrijeme pojavljivanja i grafički dizajn prema analiziranim preferencijama korisnika. Njihov utjecaj je osobito izražen kod političkog oglašavanja gdje se mikrociljanje koristi u političkim kampanjama, procesima javnog zagovaranja i nastojanjima utjecanja na javno mnijenje oko nekih politički relevantnih tema (Witzleb i Paterson, 2021, vidi i Cerovac, 2023). Poznati primjer upotrebe mikrociljanja kako bi se utjecalo na izborne procese i manipuliralo rasuđivanjem građana predstavlja kontroverza vezana uz konzultantsku tvrtku Cambridge Analytica koja je preko društvene mreže Facebook došla do osobnih podataka oko 87 milijuna korisnika (od kojih golema većina nije bila svjesna da tvrtka raspolaže njihovim podacima), te je koristeći prikupljene podatke i algoritme mikrociljanja utjecala na izborne procese poput Američkih predsjedničkih izbora 2016. godine i Referenduma o Brexitu (Kang i Frenkel, 2018; Webb i sur., 2021).

Mikrociljanje pogonjeno algoritmima umjetne inteligencije predstavlja učinkovit alat za utjecanje na promišljanje pojedinaca (Bonicalzi i sur., 2023). Koristeći se različitim strategijama, uključujući eksploataciju strahova ili predrasuda korisnika, pojedince se navodi na prosuđivanje utemeljeno na emocijama (a ne na praktičnom umu) te se otežava kritičko promišljanje kroz preopterećivanje kognitivnih sposobnosti korisnika oglasima istog tipa, a sve s ciljem stvaranja epistemičkog mjehura koji će pojedinca izolirati od drugih izvora informacija. Budući da je cijeli proces usmjeren na manipulaciju rasuđivanjem pojedinaca, korištenje algoritama umjetne inteligencije kako bi se kroz mikrociljanje distribuiralo plaćene oglase predstavlja jasan slučaj napada na autonomiju volje pojedinca.

(iv) Veliki jezični modeli

Sve češća upotreba velikih jezičnih modela pogonjenih algoritmima umjetne inteligencije može predstavljati još jedan oblik ugroze negativne slobode pojedinaca. Prošlogodišnja istraživanja pokazuju kako 89 % učenika i studenata u SAD-u koristi ChatGPT za obavljanje školskih i fakultetskih obaveza, a 53 % ga koristi za pisanje domaćih i seminarskih radova (Yu, 2023). Uz to, sve veći broj ljudi koristi velike jezične modele za prikupljanje informacija vezanih uz zdravlje, financije ili politiku. Iako sposobnost ovih algoritama da u vrlo kratkom vremenu analiziraju i ususstave velike količine podataka može biti izrazito korisna u rješavanju zadataka, ističu se (barem) dva ključna problema.

Prvo, ovi modeli su trenirani na velikim količinama podataka, no krajnji korisnici nisu upoznati s izvorima informacija iz kojih veliki jezični modeli izvlače informacije. Sadržaj koji kreiraju ovi modeli tako može isključiti neke važne perspektive ili par promovirati neke postojeće pristranosti (ovisno o podacima na kojima je model treniran). Primjerice, uočeno je kako je ChatGPT sustavno pristran prema liberalnim svjetonazorima te gaji predrasude prema konzervativcima (McGee, 2023). Osim toga, budući da modeli uče iz postojećih materijala, koji su i sami često pristrani i oblikovani društvenim uvjetima i epistemičkim okolnostima u kojima su nastali, rezultati mogu perpetuirati ili pojačavati epistemičke nepravde u društvu nepravedno marginalizirajući perspektive i stavove potlačenih društvenih skupina (Kay i sur., 2024, vidi također Samaržija i Cerovac, 2021).

Drugo, oslanjanje na velike jezične modele za rješavanje zadataka uvelike smanjuje kritički odnos prema primljenim informacijama. Korisnici prebacuju velik dio kognitivnog posla i racionalnog promišljanja na algoritme umjetne inteligencije, što pak kod korisnika (a pogotovo kod djece) dovodi do gubitka nekih važnih vještina vezanih uz razumijevanje konteksta rasprave, inovativno razmišljanje, pa čak i interpersonalnu komunikaciju (Kasneci i sur., 2023). U velikom broju slučajeva

(preko 70 %) korisnici velikih jezičnih modela kao što je ChatGPT uopće ne provjeravaju i ne sumnjaju u rezultate koje im model daje (Xu, Feng i Chen, 2023).

Oba problema upućuju na ugrožavanje negativne slobode pojedinca. U prvom slučaju oslanjanje na velike jezične modele čini pojedince podložnima manipulaciji kroz često pristrane algoritme koji analiziraju velike količine podataka i korisnicima pružaju konačne (a često pristrane) informacije. U drugom slučaju, riječ je o gubitku vještina kritičkog mišljenja u kojem se korisnici u prevelikoj mjeri i bez zadrški oslanjaju na umjetnu inteligenciju. Više nije problem usude li se pojedinci koristiti vlastite moći rasuđivanja, već imaju li motivaciju upustiti se u kritičko promišljanje kada im veliki jezični model nudi jednostavne i uvjerljive odgovore (koji su često prilagođeni da odgovaraju interesima, ali i pristranostima samih korisnika). Sve veće oslanjanje na velike jezične modele, pogotovo kod djece koja još nisu usvojila vještine kritičkog promišljanja, upućuje na opasnost koju ove digitalne tehnologije predstavljaju negativnoj slobodi pojedinaca.

Gubitak neovisnosti u mišljenju, suđenju i djelovanju svakako predstavlja potencijalni društveni problem koji može naštetiti kvaliteti političkih odluka (Cerovac, 2020; Coeckelbergh, 2023), no ostaje pitanje tko je kriv za ovaj gubitak. Prema strogom čitanju Kanta, krivnja leži na samim pojedincima koji se oslanjaju na digitalne tehnologije, bilo da je riječ o informiranju kroz društvene mreže i internetske tražilice ili korištenju velikih jezičnih modela za rješavanje praktičnih problema. Ipak, za potrebe ovog rada fokus neće biti na strogom čitanju Kanta, već na Kantom inspiriranom čitanju koje polazi od pretpostavke kako je nepoželjno (i, ako želimo funkcionirati u suvremenom svijetu, postaje gotovo nemoguće) odustati od upotrebe digitalnih tehnologija kako bismo očuvali vlastitu samostalnost u promišljanju, suđenju i djelovanju. Odbacivanje produkata tehnološkog razvoja na sličan način kao što su to radili ludisti početkom 19. stoljeća ne samo da onemogućava funkcioniranje u suvremenom svijetu, nego i uskraćuje pristup pozitivnim učincima koje digitalne tehnologije mogu imati na naše moralne i epistemičke prakse. Iako nije nemoguće potpuno odustati od oslanjanja na digitalne tehnologije pogonjene algoritmima umjetne inteligencije, pristup ovog rada ne poziva na vraćanje u prošlo stanje u kojem smo se manje oslanjali na tehnologiju, već zagovara pristup usmjeren na promišljanje o umjerenoj i mudroj regulaciji ovih tehnologija kako bi se umanjio negativni učinak koji imaju na negativnu slobodu pojedinaca.

REGULACIJA DIGITALNIH TEHNOLOGIJA KAO POKUŠAJ OČUVANJA NEGATIVNE SLOBODE GRAĐANA?

Dosadašnja je analiza pokazala da digitalne tehnologije pogonjene algoritmima umjetne inteligencije mogu doprinijeti ugrožavanju negativne slobode, koja je važna primjenu moralne autonomije građana na konkretna moralna pitanja. Osim što se često koriste za manipulaciju rasuđivanja građana i za kontroliranje javnog mnijenja kroz neke već poznate mehanizme (distribucija lažnih vijesti, političko oglašavanje i mikrociljanje), ove tehnologije oblikuju epistemičko okruženje u kojem je, čak i u slučajevima kada nema internacionalne manipulacije inicirane od strane drugih ljudi, sve teže služiti se vlastitim umom (korištenje velikih jezičnih modela, sustavi preporuka i povezani epistemički baloni). Preostaje vidjeti slijedi li, prema jednom od čitanja Kantove misli (utemeljenom na interpretaciji Thomasa Hilla), iz ove ugroze i dužnost da se uspostavit regulacija digitalnih tehnologija koja će zaštititi ili čak promicati autonomiju pojedinaca.

Iako Kant ne govori izravno o dužnostima za stvaranje sigurnog i poticajnog epistemičkog okruženja, zahvaća brojne povezane teme iz kojih se može izvesti i njegova podrška regulaciji alata i tehnologija koje ugrožavaju autonomiju volje. Pišući o autonomiji, Kant smatra kako pojedinac ima dužnosti prema sebi i dužnosti prema drugima. U prvom slučaju, dužnost je čuvati vlastitu autonomiju kroz unaprjeđivanje vlastitih racionalnih sposobnosti te izbjegavanje epistemičkih okruženja koja će dovesti do samonametnutog neznanja[8]. U drugom slučaju, dužnost je ne ograničavati ili ugrožavati autonomiju volje drugih, što će uključivati i zabranu manipuliranja drugima te ograničavanje upotrebe alata, tehnologija ili procedura koje će druge ostaviti u neznanju ili će štetno djelovati na njihove epistemičke prakse i sposobnost da slobodno koriste vlastiti um (Kant, 2002, vidi i Hill, 1992). Na ovo upućuju i brojni suvremeni interpreti Kantove misli kada pišu kako su „najviše dužnosti ljudskih bića osigurati da drugi ljudi prakticiraju neometanu autonomiju volje te da dobivaju poštovanje koje njihovo dostojanstvo zaslužuje, kao i brinuti za blagostanje drugih i tretirati ih s poštovanjem“ (Faroso, 2019, str. 81)[9]. Na sličan se način može interpretirati i Kantov poznati poziv ljudima da se usude koristiti vlastitim umom (Kant, 2010): uz jasnu osobnu primjenu ovog poziva, Kant govori i o njegovoj društvenoj primjeni. U tom slučaju, poziv traži osiguravanje društvenih

[8] Ova dužnost proizlazi iz naše dužnosti poštovanja dostojanstva u nama samima i u svakog drugoj osobi.

[9] Slično zagovara i Hill koji piše kako „opseg naših izbora može biti nelegalno ograničen na više načina – kroz fizičku silu, prijetnje i prisilu, prevare i manipulaciju, kao i kroz opresivne ideologije koje oštećuju našu sposobnost da racionalnu samo-vladavinu“. Uzevši ovo u obzir, Hill nastavlja kako „imamo snažne razloge (kao racionalni moralni zakonodavci) uspostaviti i održavati principe koji brane pravo svake osobe da vlada svojim vlastitim životom unutar određenih granica“, što će pak tražiti da „suzbijemo nelegitimna ograničavanja opsega naših izbora“ (Hill, 2013, str. 29).

uvjeta za slobodno korištenje ljudskim umom u javnom diskursu (Cronin, 2003). Drugim riječima, traži se osiguravanje društvenih preduvjeta za stvaranje epistemičkog okruženja koje će osiguravati primjenu autonomije volje građana[10].

Dužnost za stvaranje takvog epistemičkog okruženja dijelom leži na samim pojedincima, ali u velikoj mjeri i na vladama i javnim institucijama, kojima je dužnost stvoriti uvjete u kojima njihovi građani mogu slobodno i autonomno koristiti vlastiti um (Kant, 2017). Ova dužnost bi, do određene razine, obuhvaćala i sprečavanje manipulacije kroz lažne vijesti, dezinformacije i mikrociljanje, kao i sprečavanje drugih oblika heteronomije volje koji proizlaze iz korištenja velikih jezičnih modela, izloženosti sustavima preporuka i boravka unutar epistemičkih balona. Naime, iako bi se Kant složio kako pojedinci imaju moralno pravo da ne budu žrtve manipulacije, značajno je teže postaviti istu tvrdnju u kontekstu legalnog prava (zbog problema definiranja manipulacije, kao i zbog opasnosti cenzure). Zbog toga kantovski pristup ne zagovara zabranjivanje manipulacije općenito, već se iz njegove interpretacije može zagovarati samo regulaciju određenih praksi koje se mogu karakterizirati kao manipulativne. Naime, „u svojim najgorim oblicima manipulacija predstavlja oblik krađe, a zakon treba zabranjivati krađu, kako god da se odvijala" (Sunstein, 2022, str. 1960). Bilo kakav konkretan prijedlog takve regulacije značajno bi prelazio okvire ovog rada, tako da je cilj ovog dijela samo bio pokazati da unutar Kantove filozofije postoje dobri temelji za promišljanje o regulaciji digitalnih tehnologija.

Sprječavanje štetnog utjecaja koji digitalne tehnologije pogonjene algoritmima umjetne inteligencije imaju na primjenu autonomiju građana može se sagledati unutar rasprave o sprečavanju širenja lažnih vijesti i dezinformacija na internetu. Ova rasprava obuhvaća različite metode, od edukacije građana preko odgovornosti tehnoloških kompanija do pravne regulacije (Gelfert, 2021, vidi također Cerovac i Drmić, 2023). Međutim, rasprava o lažnim vijestima samo djelomično zahvaća probleme koji se vezuju uz mikrociljanje i političko oglašavanje ili uz velike jezične modele. U ovim je slučajevima ključno razumjeti kako utjecaj koji ove tehnologije vrše na rasuđivanje pojedinaca uvelike ovisi o količini informacija o korisnicima kojima algoritmi raspolažu. Ključ učinkovitosti tvrtke Cambridge Analytica u političkom oglašavanju i utjecanju na rezultate izbora nije bio u kvaliteti samog programskog koda, već u količini informacija o korisnicima koje je ova tvrtka uspjela nelegalno

[10] Thomas Hill ističe kako imamo razloge „ne samo štititi, razvijati i prakticirati naše vlastite sposobnosti za racionalnu autonomiju unutar granica koje primjereno poštuju druge, već i razvijati i podupirati društvene institucije i odnose koji promiču ovu vrijednost za sve, na primjer, kroz škole, organizacije građana, standarde etičkog novinarstva, reformu zatvora i slično". Nadodaje kako „imamo razloga brinuti o mogućnostima i resursima svake osobe da živi efektivno u kontroli nad svojim vlastitim životom, no možemo promicati tu vrijednost samo pod principima koji su pravični prema svima te poštuju moralna ograničenja oko dozvoljenih sredstava" (Hill, 2013, str. 28).

prikupiti. Prema tome, značajan napredak u uklanjanju štetnog utjecaja koji algoritmi umjetne inteligencije imaju na autonomiju pojedinaca može se postići reguliranjem vrste i tipa informacija koje kompanije mogu prikupljati o svojim korisnicima, kao i reguliranjem njihove politike upravljanja tim podacima (Zarsky, 2019) i razinom personalizacije sadržaja kroz sustave preporuka koje digitalne platforme koriste (Susser i sur., 2019).

Dok filozofska literatura nudi neke generalne smjernice za društveno-političko nastojanje zaštite autonomije pojedinaca, za izradu konkretnih rješenja potreban je obuhvatan interdisciplinarni pristup. Filozofska rasprava o autonomiji volje, uključujući i Kantov pristup koji nam ne objašnjava samo kada je prakticiranje autonomije ugroženo, već nam daje i čvrsto utemeljenje njezinog značaja, svakako predstavlja bitan element takvog (sve nužnijeg) interdisciplinarnog pristupa.

Literatura

Allcott, H. i Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.

Baccarini, E. i Prijić-Samaržija, S. (2007). *Praktična etika: Ogledi iz liberalnoga pristupa nekim problemima praktične etike*. Zagreb: Hrvatsko filozofsko društvo.

Barbu, O. (2014). Advertising, Microtargeting and Social Media. *Procedia - Social and Behavioral Sciences, 163*, 44 – 49.

Baron, J. (2016). A Welfarist Approach to Manipulation. *Journal of Marketing Behavior, 1*(3-4): 283-291.

Bonicalzi, S., De Caro, M. i Giovanola, B. (2023). Artificial Intelligence and Autonomy: On the Ethical Dimension of Recommender Systems. *Topoi, 42*(3), 1-14.

Cerovac, I. (2020). *Epistemic Democracy and Political Legitimacy*. Cham: Palgrave Macmillan.

Cerovac, I. (2022). *John Stuart Mill and Epistemic Democracy*. Lanham: Lexington Books.

Cerovac, I. (2023). Economic Inequalities and Epistemic Democracy, u H. Samaržija i Q. Cassam (ur.), *The Epistemology of Democracy*, str. 250-269. London: Routledge.

Cerovac, I. i Drmić, H. (2023). Fake News and the Capability Approach: How Disinformation Impairs Personal Health. *Prolegomena, 22*(1), 27-51.

Christiano, T. (2010). The Uneasy Relationship Between Democracy and Capital. *Social Philosophy and Policy, 27*(1), 195-217.

Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics, 3*, 1341–1350.

Consentino, G. (2020). *Social Media and the Post-Truth World Order: The Global Dynamics of Disinformation*. Cham: Palgrave Pivot.

Cronin, C. (2003). Kant's Politics of Enlightenment. *Journal of the History of Philosophy, 41*(1): 51-80.

Ecker, U., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L., Brashier, N., Kendeou, P., Vraga, E. i Amazeen, M. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology, 1,* 13-29. 10.1038/s44159-021-00006-y.

Endert, J. (2024). Generative AI Is the Ultimate Disinformation Amplifier. *Akademie* (pristup: 26.03.2024). Dostupno na: https://p.dw.com/p/4doP4

Eterović, I. (2017). *Kant i bioetika*. Zagreb: Pergamena.

Fasoro, S. A. (2019). *Kant on Human Dignity: Autonomy, Humanity, and Human Rights*. *Kantian Journal, 38*(1), 81-98.

Ferrara, E. (2020). Bots, Elections, and Social Media: A Brief Overview, u K. Shu, S. Wang, D. Lee i H. Liu, (ur.). *Disinformation, Misinformation, and Fake News in Social Media*. Cham: Springer.

Ferretti, M. P. (2018). *The Public Perspective: Public Justification and the Ethics of Belief*. Lanham: Rowman and Littlefield.

Frierson, P. R. (2005). The Moral Importance of Politeness in Kant's Anthropology. *Kantian Review, 9*, 105-127.

Gelfert, A. (2021). What is fake news?, u M. Hannon i J. de Ridder (ur.), *The Routledge Handbook of Political Epistemology*, str. 171-180. London, Routledge.

Hameleers, M., van der Meer, T. i Dobber, T. (2024). Distorting the Truth versus Blatant Lies: The Effects of Different Degrees of Deception in Domestic and Foreign Political Deepfakes. *Computers in Human Behavior, 152*, 108096.

Hao, K. (2019). Why AI is a threat to democracy – and what we can do to stop it. *MIT Technology Review*. Dostupno na: https://www.technologyreview.com/2019/02/26/66043/why-ai-is-a-threat-to-democracyand-what-we-can-do-to-stop-it/

Hill, T. E. (1989). The Kantian Conception of Autonomy, u J. Christman (ur.), *The Inner Citadel: Essays on Individual Autonomy*, str. 91-108. New York: Oxford University Press.

Hill, T. E. (1992). *Dignity and Practical Reason in Kant's Moral Theory*. Ithaca: Cornell University Press

Hill, T. E. (2013). Kantian autonomy and contemporary ideas of autonomy, u O. Sensen (ur.), *Kant on Moral Autonomy*, 15-31. Cambridge: Cambridge University Press.

Kang, C. i Frenkel, S. (2018). Facebook Says Cambridge Analytica Harvested Data of Up to 87 Million Users. *The New York Times*. Dostupno na: https://www.nytimes.com/2018/04/04/technology/mark-zuckerberg-testify-congress.html

Kant, I. (2010). *An Answer to the Question: 'What Is Enlightenment?'*. New York: Penguin Books.

Kant, I. (2002). *Groundwork of the Metaphysics of Morals*. Oxford: Oxford University Press.

Kant, I. (2015). *Critique of Practical Reason*, 2nd edition. Cambridge: Cambridge University Press.

Kant, I. (2017). *The Metaphysics of Morals*, Revised Edition. Cambridge: Cambridge University Press.

Kasneci, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Kay, J., Kasirzadeh, A. i Mohamed, S. (2024). Epistemic Injustice in Generative AI. *arXiv 2408*,11441.

Kiri Gunn, H. (2021). Filter bubbles, echo chambers, online communities, u M. Hannon i M. de Ridder (ur.), *The Routledge Handbook of Political Epistemology*, str. 192–202. London: Routledge.

Korsgaard, C. M. (2007). What's Wrong With Lying?, u J. E. Adler i C. Z. Elgin (ur.), *Philosophical Inquiry: Classic and Contemporary Readings*. Indianapolis: Hackett Publishing Company.

La Morgia, M., Mei, A., Mongardini, A. M. i Wu, J. (2021). Uncovering the Dark Side of Telegram: Fakes, Clones, Scams, and Conspiracy Movements. *arXiv 2111*,13530.

Margalit, A. (2016). Autonomy: Errors and Manipulation. *Jerusalem Review of Legal Studies*, *14*(1): 102–112.

McGee, R. W. (2023). Is Chat GPT Biased Against Conservatives? An Empirical Study. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4359405

McIntyre, L. C. (2018). *Post-Truth*. Cambridge: MIT Press.

McKay, S. i Tenove, C. (2021). Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly*, *74*(3), 703-717.

Menczer, F. i Hills, T. (2020). Information Overload Helps Fake News Spread, and Social Media Knows It. *Scientific American – Springer Nature America*. Dostupno na: https://www.scientificamerican.com/article/information-overload-helps-fake-news-spread-and-social-media-knows-it/

Mill, J. S. (1977). Considerations on representative government, u J. M. Robson (ur.), *Collected Works of John Stuart Mill*, Vol. 19, str. 371–578. Toronto: University of Toronto Press.

Mill, J. S. (2020). *O slobodi*. Zagreb: Jesenski i Turk.

Moles, A. (2007). *Autonomy, Freedom of Speech and Mental Contamination*. Neobjavljena doktorska disertacija. Warwick: University of Warwick.

Nguyen, T. C. (2020). Echo chambers and epistemic bubbles. *Episteme, 17*(2): 141–161.

Pariser, E. (2011). *The Filer Bubble: How the New Personalized Web is Changing What We Read and How We Think.* London: Penguin Books.

PEW Research Center. (2024). News Platform Fact Sheet. (Pristup: 17.9.2024). Dostupno na: https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/

Piper, M. (2024). "Autonomy: Normative", u *The Internet Encyclopedia of Philosophy*. Dostupno na: https://iep.utm.edu/normative-autonomy/

Pollock, J. (2024). Epistemic Bubbles and Contextual Discordance. *Philosophy, 99*(3): 437-459.

O'Neill, O. (2014). *Acting on Principle: An Essay on Kantian Ethics*, 2nd edition. Cambridge: Cambridge University Press.

Quong, J. (2010). *Liberalism without Perfection*. Oxford: Oxford University Press.

Rapp, D. N. (2016). The Consequences of Reding Inaccurate Information. *Current Directions in Psychological Science, 25*(4), 281-285.

Rawls, J. (2005). *Political Liberalism*. New York: Columbia University Press.

Reath, A. (2006). *Agency and Autonomy in Kant's Moral Theory*. Oxford: Oxford University Press.

Rhodes, S. C. (2022). Filter bubbles, echo chambers, and fake news: how social media conditions individuals to be less critical of political misinformation. *Political Communication, 39*(1): 1–22.

Rini, R. (2017). Fake News and Partisan Epistemology. *Kennedy Institute of Ethics Journal, 27*(2): E43-E64.

Sahebi, S. i Formosa, P. (2022). Social Media and its Negative Impacts on Autonomy. *Philosophy and Technology, 35*, 70.

Simchon, A., Edwards, M. i Lewandowsky, S. (2014).The Persuasive Effects of Political Microtargeting in the Age of Generative Artificial Intelligence, *PNAS Nexus, 3*(2), 1-5.

Singer, D. J., Grim, P., Bramson, A., Holman, B., Jung, J. i Berger, W. J. (2021). Epistemic networks and polarization. U M. Hannon i J. de Ridder (ur.). *The Routledge Handbook of Political Epistemology*, str. 133-144. London: Routledge.

Spitale, G., Biller-Andorno, N. i Germani, F. (2023). AI Model GPT-3 (Dis)Informs Us Better than Humans. *Science Advances, 9*(26), str. 1-9.

Sunstein, C. R. (2009). *Going to Extremes: How Like Minds Unite and Divide*. Oxford: Oxford University Press.

Sunstein, C. R. (2022). Manipulation as Theft. *Journal of European Public Policy, 29*(12): 1959–1969.

Susser, D., Roessler, B. i Nissenbaum, H. (2019). Technology, Autonomy, and Manipulation. *Internet Policy Review, 8*(2), 1-22.

Tufekci, Z. (2014). Engineering the Public: Big Data, Surveillance and Computational Politics. *First Monday, 19*, 1-39.

Webb, M., Dowling, M. i Farina, M. (2021). *Understanding Mass Influence: A Case Study of Cambridge Analytica as a Contemporary Mass Influence Campaign*. Adelaide: University of Adelaide.

Witzleb, N. i Paterson, M. (2021). Micro-targeting in Political Campaigns: Political Promise and Democratic Risk. U U. Kohl i J. Eisler (ur.). *Data-Driven Personalisation in Markets, Politics and Law*, str. 223-240. Cambridge: Cambridge University Press.

Xu, R., Feng, Y., i Chen, H. (2023). ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience. *arXiv* 2307:01135.

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology, 14*. doi.org/10.3389/fpsyg.2023.1181712

Zarsky, T. Z. (2019). Privacy and Manipulation in the Digital Age. *Theoretical Inquiries in Law, 20*(1), 157–188.

Zhang, Q., Lu, J. i Jin, Y. (2021). Artificial Intelligence in Recommender Systems. *Complex and Intelligent Systems,* 7, 439–457.

Zhang, D. (2020). China's Digital Nationalism: Search Engines and Online Encyclopedias. *The Journal of Communication and Media Studies, 5*(2), 1-19.

Zimdars, M. i McLeod, K. (2020). Fake News: Understanding Media and Misinformation in the Digital Age. Cambridge: MIT Press.

Zinkin, M. (2024). *Depth: A Kantian Account of Reason*. New York: Oxford Academic.

Fake News, Digital Technologies, and the Erosion of the Realisation of Individual Autonomy in Light of Kantian Ethics

SUMMARY

Digital technologies have been radically transforming the epistemic environment of citizens. By creating the way citizens collect information, communicate, or make decisions, digital technologies form new epistemic practices, which provide space for some old as well as new forms of manipulation. This paper begins with an analysis of Kant's conception of the autonomy of the will and shows how and in what cases this autonomy can be threatened. The paper continues by providing insights into how digital technologies can threaten the autonomy of citizens. It analyses the ability of AI algorithms to spread fake news and political propaganda through microtargeting and recommender systems. Furthermore, the paper considers the harmful influence of these technologies on the epistemic practices of citizens, emphasising the AI algorithm's tendency to create epistemic bubbles or overreliance on large language models, thereby weakening individual's judgement ability. Finally, some regulation models of these technologies are considered, and the Kantian stronghold to justify such practices is highlighted.

Keywords: recommender systems, microtargeting, large language models, fake news, epistemic bubbles, heteronomy of will.