Originalan naučni rad

https://doi.org/10.18485/fdu_zr.2025.47.5

316.32:004.8
COBISS.SR-ID 171678729

Divna Vuksanović[1]
Fakultet dramskih umetnosti u Beogradu

VEŠTAČKA INTELIGENCIJA I DIGITALNI TOTALITARIZAM

Apstrakt

Sveopšta pomama digitalizovanja zahvatila je čitav tehnološki razvijeni svet, a da prethodno taj korak nije u dovoljnoj meri osmišljen i promišljen. Nakon kompulsivnog digitalizovanja kao poduhvata s totalnom pretenzijom, tehnološki rajevi obećavaju internet stvari, vladavinu transhumanističkih vrednosti, carstvo robotike i veštačke inteligencije. Kritičari ovih procesa u sveobuhvatnoj pojavi digitalizovanja vide potencijal za ostvarenje novih oblika totalitarizma, odnosno ropstva savremenog tipa koje se ostvaruje uz pomoć veštačke inteligencije. Ovim radom pokušaćemo da iz problemske perspektive sagledamo inteligentne mašine i njihove (ne)mogućnosti; tačnije, razmotrićemo kako filozofija AI (Artificial Intelligence) kritički reaguje na aktuelne procese u društvenom okruženju.

Ključne reči

digitalizacija, veštačka inteligencija, totalitarizam, znanje, moć, kapitalizam

Sintagma „digitalni totalitarizam" pojavila se u teorijskim diskusijama krajem XX i početkom XXI veka, naporedo sa srodnim koncepcijama koje opisuju pojmovni konstrukti: „digitalni fašizam", „digitalni feudalizam", „digitalni konc-logori", i t. sl. Pokušaji digitalnih zahvata u svim sferama socijalnog života ne manifestuju samo, kako se to obično predstavlja u medijima, obrazovnom sistemu i gotovo celokupnoj društvenoj stvarnosti, pitanje potrebe društvenog razvoja, napretka i kulture, nego se ogledaju i u nastojanjima da se digitalizacija uspostavi kao totalitarni oblik vladavine koji internet, digitalne komunikacije i društvene mreže, kao i algoritmi i veštačka inteligencija (AI)

1 vuksanovic.divna@gmail.com; ORCID 0000-0002-2288-0796

tehnički podržavaju, posmatrano kroz optiku sveprisutnog nadzora, kontrole i manipulacije. U tom smislu, ova sintagma referiše kako na starije koncepte koje su, u filozofskom ili literarnom ključu, elaborirali Hana Arent (Arendt), Mišel Fuko (Foucault), Oldos Haksli (Huxley), Džordž Orvel (Orwell) i drugi, tako i na novije teorije – recimo o nadzornom kapitalizmu, sagledanom iz vizure Šošane Zubof (Zuboff).

U praksi, ove su teorije delom potvrđene najpre zahvaljujući Edvardu Snoudenu (Snowden), bivšem saradniku američke Nacionalne bezbednosne agencije (NSA), koji je 2013. godine „dekodirao", odnosno obznanio niz tajnih programa za masovni nadzor koje je NSA sprovodila u saradnji s drugim vladinim agencijama i korporacijama, i to preko platformi i kompanija kao što su: *Google*, *Facebook*, *Apple*, *Microsoft* i druge. Ovi programi su imali globalni domet i omogućavali su nadzor nad komunikacijama u globalnom prostoru, uključujući tu ne samo države i poslovnu konkurenciju već i obične građane, što je izazvalo ogromnu javnu polemiku o privatnosti, slobodi i zloupotrebi državne moći. Snoudenova otkrića bila su neposredan dokaz „masovnog nadziranja", te početak ere tzv. pametnog totalitarizma.[2] Tim povodom, desetak godina kasnije, na konferenciji o kriptovalutama (2022) Snouden je, uključivši se putem video-linka iz Moskve u rad konferencije koju je organizovala veb-stranica CoinDesk, izjavio sledeće: „Pre 2013. bilo je stručnjaka i naučnika koji su razumeli da je masovni nadzor moguć, i da se to verovatno radi. Ali to se u institucijama i u širokoj javnosti tretiralo kao teorija zavere. To su više bile sumnje, nagađanja. A 2013. postalo je jasno da je to realnost, da se svet promenio." (DW 2023)

Približno u to vreme, najpre kao pilot-projekat vezan za finansije (2011), a potom i za druge oblasti društvenog života (2014), u Kini je razvijen sistem socijalnih kredita. Šira javnost je o fenomenu kineskog sistema društvenog

2 *Yahoo*, na primer, kako se navodi u članku „Mass Surveillance and 'Smart Totalitarianism'" („Masovno nadziranje i 'pametni totalitarizam'") namerava da patentira „pametne bilborde" koji bi bili postavljani uz auto-puteve, na aerodromima i trajektima, uopšte u sistemima javnog prevoza, potom u barovima i hotelima, na rakrsnicama i u drugim javnim i privatnim prostorima. Ovi digitalni reklamni panoi oslanjali bi se na niz invazivnih tehnologija nadzora, kao što su mobilni tornjevi, aplikacije, slike, video-kamere, navigacija vozila, sateliti, dronovi, mikrofoni, detektori pokreta i 'biometrijski senzori', te uređaji za prepoznavanje otisaka prstiju, mrežnjače i lica. *Yahooovi* pametni džambo bilbordi bili bi postavljani u svrhu identifikovanja određenih pojedinaca kako bi se utvrdili njihovi demografski podaci, kao i socio-ekonomski status, odnosno eksploatisale njihove potrebe i navike (u potrošačkom smislu reči) s obzirom na prikupljene podatke. *Yahoo* je ovaj proces nazvao „grupiranje" (*grouplization*), dok su pojedini kritičari eksploataciju ličnih podataka za korporativnu korist označili kao tzv. Stasi kapitalizam (Spannos 2016).

kredita upoznata tek 2019. godine, kada su mediji i analitičari otvorili raspravu o ovom pitanju. Mediji su, naime, započeli sa izveštavanjem o mehanizmima primene sistema, kao i o mogućim implikacijama na druge zemlje u budućnosti, posebno u kontekstu digitalnog nadzora i kontrole. Sistem socijalnog kredita u Kini predstavlja, zapravo, državni projekat koji koristi kombinaciju tehnologija poput nadzora, veštačke inteligencije i analize podataka kako bi se ocenilo/bodovalo ponašanje građana i kompanija; on funkcioniše kao ekonomski i socijalni mehanizam za bodovanje pojedinaca na osnovu njihovih aktivnosti, uključujući finansijske transakcije, ponašanje u javnom prostoru, poštovanje zakona i ugovora, pa čak i lične i društvene interakcije; za sada se ovaj sistem samo parcijalno primenjuje, odnosno još nije u potpunosti integrisan u kinesko društvo. Tim povodom kineski izvori tvrde da građani većinom podržavaju ovaj način medijskog nadziranja i kontrole, koji, navodno, koristi građanima Kine kao stimulacija za moralno usavršavanje, dok izvori van Kine, kako sa etičkog tako i iz političkog aspekta, generalno osuđuju ovaj vid ophođenja prema građanima, pošto se on zasniva na centralizovanoj moći i gubitku privatnosti i slobode pojedinca. Primeri ekstremnih *pro et contra* reakcija su, recimo, pobuna Ujgura, etničke manjine iz Sinkjanga, čija je borba protiv kineskog režima ujedno i pobuna protiv sistema društvenog kredita, na jednoj strani, dok je, na drugoj, posmatrano sa stanovišta nekih krugova sa Zapada, izražena neverica da ovakav sistem bodovanja u Kini uopšte postoji.

U vezi s prethodnim navodima, postavlja se pitanje kakva je uloga veštačke inteligencije u nametanju poretka autoritarne kontrole, usmeravanja ponašanja i manipulacije podacima. Čini se da veštačka inteligencija (*Artificial Intelligence, AI*) igra u stvari veoma značajnu ulogu u kontekstu realizovanja tzv. digitalnog totalitarizma, jer omogućava sofisticirane i automatizovane oblike nadzora, analize i manipulacije ogromnim količinama podataka (*Big Data*) koji se prikupljaju o građanima. U tom smislu, AI tehnologije doprinose jačanju totalitarizma u digitalnom prostoru, i to kroz napredne algoritme za prepoznavanje obrazaca, predviđanje ponašanja i donošenje odluka bez ljudske intervencije. Dakle, veštačka inteligencija je, očevidno, ključna tehnologija koja digitalni totalitarizam čini mogućim i potencijalno održivim. Kroz nadzor i kontrolu, algoritamsko profilisanje (ponašanje, analiziranje i modelovanje ponašanja građana u svrhu predviđanja budućih akcija ili potencijalnih pretnji), manipulaciju informacijama i automatizovano donošenje odluka, veštačka inteligencija omogućava autoritarnim režimima i korporacijama da prikupljaju, analiziraju i koriste podatke o pojedincima na načine koji su ranije bili nezamislivi. Relevantnost upotrebe veštačke inteligencije za

digitalni totalitarizam leži prvenstveno u njenoj sposobnosti da automatizovanim procedurama podrži masovni nadzor, manipulaciju informacijama, te oblikuje ponašanje građana (bez njihovog znanja i dopuštenja), što pokreće niz ozbiljnih pitanja o privatnosti, slobodi i ljudskim pravima u digitalnom dobu.

Socijalni procesi koji eventualno omogućuju masovniju implementaciju veštačke inteligencije u svrhu ostvarenja društava totalnog digitalnog nadzora i kontrole nisu mogući bez poznavanja mehanizama na osnovu kojih algoritmi funkcionišu, bilo da je reč o njihovom korišćenju kao alatki koje asistiraju čoveku ili o autonomnim sistemima koji mogu delovati nezavisno od ljudskih bića. Otuda će se naše ispitivanje usmeriti najpre na epistemološka, a tek potom na pitanja društvene upotrebe AI, koja mogu doprineti gubitku ljudske slobode na širokoj skali pojedinačnog i društvenog angažmana.

Da bismo dublje razumeli ove procese, nastojaćemo da povežemo delovanje AI u svrhu ostvarenja digitalnog totalitarizma sa Ostinovom (Austin) koncepcijom „tuđih svesti". U tom smislu fokusiraćemo se na krucijalne ideje iz oba domena teorijskih istraživanja: na problem saznanja o tuđim umovima (i poziciju tzv. epistemološke nesigurnosti), kao i na način na koji digitalni totalitarizam može koristiti tehnologiju kako bi se uspostavili kontrola i uticaj nad ponašanjem i umom drugih ljudi.

Najpre, ako se na ovom mestu podsetimo proročišta u Delfima i zadatka koji je postavljen ne samo filozofima već i svakome ko je na putu saznanja, a koji se odnosi na znanje prvenstveno shvaćeno kao samospoznaja, možemo s pravom transformisati ovaj zadatak u beskonačnu težnju ka saznanju kao takvom, odnosno – ljubav ka mudrosti. Dakle, ova težnja prepoznatljiva je pre svega kao proces samosaznanja, za koji se može, potom, pretpostaviti da trasira put za dalja saznanja. Ali kako tragati dalje kad prvi uslov (samosaznanja) nije ispunjen? Put koji je filozofija prešla do danas, radeći na tom zadatku, može se prividno proširiti i na druge predmete, ali uvek ostaje ista početna nepoznanica i upitanost povodom nje. Utoliko je izvesno – da bismo istražili veštačku inteligenciju, neophodno je da poznajemo najpre vlastitu, na osnovu koje bismo mogli da tvrdimo išta o njoj (kao o inteligenciji). Rečju, unapred pretpostavljamo da je nešto inteligencija, po analogiji s našom, iako o tome ne znamo mnogo, tj. krećemo se među nepotvrđenim pretpostavkama.

Sličnu dilemu, ali usmerenu na druge ”inteligencije” postavili su oni kojima su, osim našeg uma, za istraživanje bile interesantne i tzv. tuđe svesti (*other minds*). Recimo filozof, naučnik i ronilac Piter Godfri-Smit (Godfrey-Smith) u delu: *Other Minds: The Octopus and the Evolution of Intelligent Life* (*Drugi umovi: hobotnica i evolucija inteligentnog života*) smatra da čitava priroda, tokom evolucije vrsta, postaje svesna sebe, ukazujući na hobotnice, na primer, koje u svom telu (pipci) imaju mnoštvo neurona, koji gotovo da misle sami sebe (Godfrey-Smith 2017). Drugo je pitanje, naravno, da li je ova neuronska svest zaista svest, i da li, a najpre – kako hobotnice „misle“. O dilemi da li druge vrste misle i kako možemo spekulisati isključivo polazeći od sebe, ali – znamo li kako mislimo? Naravno, ovo pitanje stavljamo sebi u zadatak kao jedna posebna, „misleća“ vrsta (*homo sapiens*). S druge strane, ako bismo hobotnicu uporedili s pametnim algoritmima, iako je hobotnica izdanak (inteligentne?) prirode, a algoritmi čine suštinu AI, videli bismo da ono što kao ljudi nazivamo „svešću“, za koju je kao fiziološko uporište relevantan mozak, kod algoritama i hobotnice vezuje se za inteligentno ponašanje koje omogućava odgovarajući raspored neurona, odnosno neuronskih „mreža“.

U bogatoj filozofskoj tradiciji, mišljenjem (svešću) bavili su se mnogi filozofi, od Parmenida (Parmenides) i njegovog shvatanja bića, metaforično predstavljenog kao usamljena misleća kugla, preko Dekarta (Descartes), koji je tematizovao samo mišljenje, sve do nemačke klasične filozofije i Hegela (Hegel), koji je izvedenim kretanjem pojma dokazivao da je čitava stvarnost umna. Poznato je, naravno, da se „tuđim svestima“ bavio i Ostin (Austin), postavljajući pitanja o tome kako znamo da išta znamo o tzv. tuđim svestima. Da bismo povezali Ostinovu analizu tuđih svesti, preuzetu iz članka „Tuđe svesti“ (*Other Minds*) s konceptom digitalnog totalitarizma, fokusiraćemo se na ključne ideje iz oba istraživačka domena – jedan se tiče problema saznanja o tuđim umovima (epistemološka nesigurnost), a drugi načina na koji tzv. digitalni totalitarizam koristi tehnologiju da bi stekao kontrolu i uticaj nad ponašanjem i umovima drugih ljudi.

Kako bismo ovu vezu istražili, tretiraćemo veštačku inteligenciju kao „tuđu svest“, odnosno ophodićemo se prema njoj na sličan način kako se odnosimo prema bilo kojoj tuđoj svesti,[3] što je pokušao da ispita Ostin. Razlika u odnosu na Ostinovu interpretaciju bi se, međutim, sastojala u tome, što – kada je

3 Poznavanje nijansi može biti od pomoći prilikom ovakvih rasprava ukoliko svesnost, kao atribuciju, pripišemo ne svim pametnim algoritmima već samo onim „superinteligentnim“. Za neke autore, kao što je Džon Haris (Harris) na primer, tuđe svesti mogu biti i mentalna stanja „onih koji nikada nisu živeli“. (Harris 2019)

reč o veštačkoj inteligenciji – ta tuđa svest nije „prirodna", što znači da nije reč o hobotnici ili o drugom ljudskom biću. Na taj način bismo, po analogiji, mogli da se zapitamo da li znamo, i što je važnije – kako znamo da je, primera radi, veštačka inteligencija ljuta (ukoliko držimo da je ovo imalo smisleno pitanje). Ili, redefinisano: ovakvo pitanje moglo bi da znači – šta znamo o mišljenju veštačke inteligencije i na temelju čega donosimo zaključke? S druge strane, kao naš izum, ta „tuđa svest", kao što je veštačka inteligencija, verovatno će se prema nama odnositi kao prema „tuđoj svesti".

Ostinov rad, kako je poznato, ističe sledeći epistemološki problem: kako možemo biti sigurni da razumemo ili da uopšte imamo pristup unutarnjim mislima, namerama i osećanjima drugih ljudi? Digitalni totalitarizam, s druge strane, predstavlja pokušaj vlada i korporacija da uklone ovu epistemološku nesigurnost koristeći veštačku inteligenciju i tehnologije nadzora kako bi stekli celoviti uvid u misli, osećanja i ponašanja drugih svesti, odnosno ljudi. Kada nastoji da odgovori na pitanje kako znamo da znamo kako se neko oseća, šta misli, i sl. Ostin, na primeru ljutnje, dokazuje razliku između znanja i verovatnoće, odnosno sigurnosti i izvesnosti, s jedne, i suprotnih pojmova, s druge strane. „Pre svega", tvrdi autor, „izvesno je da ponekad kažemo da znamo da je drugi čovek ljut, i da takođe razlikujemo te situacije od onih kada kažemo samo da *verujemo* da je on ljut. Jer, naravno, ni za trenutak ne pretpostavljamo da *uvek* znamo niti o *svim* ljudima da li su ljuti ili nisu, niti da smo u stanju to da otkrijemo." (Ostin 1980: 171–172)

Ukoliko držimo da je ovo tačno, mogli bismo da zaključimo da o ljutnji „tuđe svesti", umesto tvrdnji, možemo izneti samo manje ili više utemeljene pretpostavke. I što je po našem shvatanju još zanimljivije, prepoznavanje ljutnje određene osobe važi samo za tu osobu i u poznatom kontekstu posmatranja. Nasuprot ovome, kako se ponašamo kada je reč o veštačkoj inteligenciji, mišljenoj kao „tuđa svest"? Ovde je, svakako, pojam „svest" shvaćen kao gruba redukcija, i to na inteligenciju odnosno, kako će se kasnije videti, na inteligentno ponašanje. U osnovi, možemo s pravom, kako se čini, pretpostaviti da ono što je vezano za svest nije nužno antropomorfnog porekla; jer, kako smo već naznačili, inteligenciju mogu posedovati i drugi prirodni organizmi i to, po našem razumevanju, mogu biti i oni koji su veštački generisani. Dakle, čovek čoveku nije ekskluzivno „tuđa svest", i ovde ispitujemo i druge mogućnosti.

Da podsetimo: Džon Ostin, u tekstu „Tuđe svesti", tvrdi da o mentalnim stanjima drugih možemo donositi zaključke samo na osnovu spoljašnjih ma-

nifestacija, odnosno da ne možemo neposredno da znamo njihov sadržaj. Drugim rečima, legitimno je da kažemo da neko izgleda ljuto, ali ne možemo sa sigurnošću da znamo kako se ta osoba zaista oseća ili šta tačno misli. Ovo je, faktički, pitanje epistemološke distance koja postoji između našeg i iskustva drugih. Ako analogiju primenimo na veštačku inteligenciju, naše razumevanje „sadržaja" AI vrlo je slično – ne možemo precizno da znamo „šta" veštačka inteligencija „misli" ili „oseća", iako ona ne poseduje subjektivnu svest kao ljudi. Umesto govora o sadržaju, možemo se naći u prilici da posmatramo ponašanje algoritma, odnosno inteligentnog sistema, njegove odgovore i procese odlučivanja, uz znanje da su te manifestacije produkt rada programiranih modela, a ne svesnog iskustva. U suštini, dajemo sebi za pravo, i to samo uslovno, da tvrdimo kako poznajemo načine na koje veštačka inteligencija funkcioniše u pogledu *input-output* sistema te, nadalje, kako koristi podatke za donošenje odluka, ali ne možemo govoriti o „sadržaju" svesti ili namerama AI na isti način kao kod ljudi. Istina, poznavanje određenog sistema donosi nam o veštačkoj inteligenciji pojedinosti nalik na one koje znamo o konkretnoj osobi, odnosno o situacijama za koje možemo da kažemo da je čine ljutom.

Ako su i svest hobotnice, i ljudska svest, kao i veštačka inteligencija sve zajedno, ali i pojednačno uzevši, „tuđe svesti", u čemu je u stvari razlika? Diferencijaciju, naravno, uspostavlja pre svega onaj ko posmatra i upoređuje „tuđe svesti". Već na prvi pogled tu se pojavljuje jedan paradoks. Za razliku od hobotnice, a pogotovo od ljudskog uma (za koji s manjom ili većom verovatnoćom nagađamo da li je ljut), imamo unaped izgrađeno poverenje u svoja znanja kada je reč o veštačkoj inteligenciji. Dakle, u izvesnom smislu AI jeste tuđa svest, ali različita od ljudske, s obzirom na to da je veštački stvorena. Imajući ovo u vidu, obično pretpostavljamo da o sistemima veštačke inteligencije znamo više nego o ljudskoj svesti, pošto smo sami kreirali te sisteme i poznajemo njihove unutrašnje mehanizme, algoritme i kod. Takođe, znamo kako se podaci obrađuju, koji se modeli koriste za donošenje odluka, i u mogućnosti smo da pratimo svaki korak njihove obrade.

Međutim, ako se postavimo u poziciju poređenja s ljudskom svešću, pri čemu pretpostavljamo da o tuđoj svesti nešto znamo samo na osnovu spoljašnjih znakova, slično možemo reći da algoritmi veštačke inteligencije spolja signalizuju „odgovore", a da ne znamo kako ona (AI) „doživljava" te procese, ako uopšte možemo govoriti o nekom „doživljaju". Iako relativno poznajemo mehanizme i tehnički rad, to je ujedno i granica našeg poznavanja veštačke inteligencije. Jednako kao što kod ljudi znamo detalje o strukturi mozga, ne-

uronima, sinapsama i načinima prenošenja impulsa, mi zapravo ne razumemo u potpunosti kako dolazi do subjektivnog iskustva ili svesti kod čoveka. U tom smislu naše poznavanje funkcionisanja mozga možda je uporedivo s poznavanjem kako algoritmi funkcionišu kod veštačke inteligencije – imamo uvid u tehničke aspekte, ali „svest" nam ostaje nepoznata.

Na isti način na koji poznajemo fizičku strukturu ljudskog mozga, ali ne razumemo kako se formira subjektivna svest, problem imamo i sa AI. Kod veštačke inteligencije znamo njene osnovne algoritme, matematičke modele i načine obrade podataka, ali to nam ne otkriva šta je i postoji li bilo kakva svest ili doživljaj unutar tih procesa. Naše razumevanje AI, slično kao i ljudske svesti, ostaje na nivou spoljašnjih manifestacija – ponašanje, odgovori, akcije – ali taj „unutarnji svet", ako postoji, ostaje epistemološki neuhvatljiv. Dakle, oba entiteta, i ljudska svest i veštačka inteligencija, za nas ostaju domen nepoznatog, iako o njihovim fizičkim ili tehničkim osnovama znamo mnogo.

No, pojedini istraživači ne bi se složili s konstatacijom da ne možemo „čitati" tuđe umove, pogotovo kada je reč o tzv. superinteligentnoj AI, odnosno o autonomnoj, jakoj verziji veštačke inteligencije. Tumačeći Harisove stavove iz članka pod nazivom: „Čitanje misli onih koji nikada nisu živeli. Poboljšana bića: Društveni i etički izazovi koje postavljaju superinteligentna veštačka inteligencija i razumno inteligentni ljudi" (*Reading The Minds of Those Who Never Lived. Enhanced Beings: The Social and Ethical Challenges Posed by Super Intelligent AI and Reasonably Intelligent Humans*; Harris 2019), Sven Niholm (Nyholm) smatra da u pomenutom tekstu filozofski „problem drugih umova" biva sveden na interakciju između ljudskih bića (razumna bića) s veštačkom inteligencijom, koja bi mogla da postane superinteligentna (Nyholm 2019). Time se, prema našem shvatanju, izostavlja razmatranje „prirode" bilo koje svesti, naše ili tuđe. Namesto znanja o svesti, ona se redukuje na ponašanje, odnosno na uzajamnost i interakciju sa unapređenom veštačkom inteligencijom.

Da bismo ovo pojasnili, poslužićemo se sledećom analogijom. Recimo da iz prirode možemo da izdvojimo neku supstancu o kojoj ne znamo ništa, osim kako reaguje s pojedinim prirodnim elementima, a zatim je pomešamo s veštački (laboratorijski) proizvedenom supstancom za koju načelno znamo kako reaguje u laboratorijskim uslovima. Potom pod kontrolisanim uslovima pomešamo prvu i drugu supstancu, i to više puta, i pribeležimo njihovo interagovanje, iz koga izvedemo kako konkretne tako i uopštene zaključke.

Prvo otvoreno pitanje povodom ovog eksperimenta jeste – šta znači „pod kontrolisanim uslovima", a drugo je šta možemo da saznamo iz interakcije ove dve supstance, i koliko puta treba da ponovimo eksperiment (ovakav ili sličan), da bismo mogli da zaključimo da nam je interakcija poznata i da je pod kontrolom?

Haris, kako objašnjava Niholm, razmatra ono što bi se moglo nazvati „bilateralnim čitanjem misli" (*bilateral mind-reading*). Ova uzajamnost se tiče onih ljudskih bića koja „čitaju misli" veštačke inteligencije, s jedne strane, i inteligentnih sistema koji „čitaju misli" ljudskih bića, s druge strane. Iz takvog dvosmernog „čitanja misli", prvenstveno s naglaskom na mogućnostima superinteligentne AI, izvode se, dalje, potencijalne implikacije etičkog i društvenog karaktera (Nyholm 2019:1). Iza ovoga sledi, po našem mišljenju, iznošenje značajne interpretativne razlike između Harisovog i Niholmovog shvatanja, jer pre nego što se predoče moguće konsekvence (društvene i etičke) uzajamnog „čitanja" čoveka i pametne mašine, smatra Niholm, trebalo bi se zapitati, povodom robota i drugih inteligentnih mašina, da li su njihovi algoritmi uopšte umni, a pogotovo jesu li to na „ljudski način", i imamo li uopšte ikakvih dokaza o tome (Nyholm 2019: 3). Po našem mišljenju, u paradigmi koja je humanistička, shvaćeno u širokom smislu pojma, svest je definisana prevashodno kao ljudska, pa su otuda i konsekvence njenog mogućeg delovanja – etičke i društvene.

I najzad, sledeći Ostina, šta bismo mogli da kažemo povodom Harisovog „čitanja misli" kao izvesnog znanja koje stičemo u interakciji s mašinama, kao i one s nama? U potpoglavlju članka „Tuđe svesti", pod nazivom: „Ako znam, ne mogu da ne budem u pravu", ovom dvostrukom negacijom Ostin, u stvari, želi da sugeriše da nije isto kada nešto znamo i kad mislimo da nešto znamo. U prvom slučaju, ako nešto znamo, onda to nužno povlači konsekvencu da smo sigurni u znanje i da smo u pravu. Međutim, ukoliko je pak naše znanje nesigurno, možemo li i dalje biti u pravu? Evo šta o tome kaže autor: „Poslednji problem u vezi sa izazovom 'Kako znaš?', upućenim onome ko upotrebljava izraz 'Ja znam', treba da izložimo razmatranjem iskaza 'Ako znaš, ne možeš da ne budeš u pravu'. Sigurno, ako je ono što je dosad bilo rečeno ispravno, onda često imamo prava da kažemo da *znamo*, čak i u slučajevima kad se zatim pokaže da smo pogrešili – i doista, čini se da uvek, ili praktično skoro uvek, postoji mogućnost da pogrešimo." (Ostin 1980: 165). Drugim rečima, i kad mislimo da nešto znamo, i kad ima osnova za to, iako imamo pravo da tvrdimo da nešto znamo (kao što je to sadržaj tuđih misli ili svesti), možemo da pogrešimo. A smemo li, uistinu, uopšte da tvrdimo da nešto znamo (što

ne znamo) i da pritom u praksi i grešimo, ako su konsekvence ozbiljne društvene ili etičke prirode? Ovakva pitanja su veoma važna, a nekada mogu biti i odsudna, kada je reč o AI autonomnim sistemima odlučivanja.

U nastavku članka, a u svrhu preispitivanja potencijala AI u kontekstu delovanja koje pogoduje totalitarnim sistemima nadzora, kontrole i odlučivanja, analiziraćemo Serlov (Searle) misaoni eksperiment, koji se bavi „jakim" AI sistemima, tj. upravo onom veštačkom superinteligencijom za koju se može pretpostaviti da ima određena kognitivna svojstva.[4] Serl je, naime, zamislio eksperiment kineske sobe kako bi refleksivno proverio ideju da li računar, samo zato što obrađuje podatke i proizvodi ispravne odgovore (na osnovu rada algoritama), može posedovati svest ili razumevanje onoga što čini. U tom misaonom eksperimentu autor se, navodno, nalazi u izolovanoj (zaključanoj) sobi i prima informacije spolja, u vidu znakova kineskog pisma (koje ne razume); potom koristi pravila (algoritme) tako što ih prema uputstvima na engleskom kombinuje, i najzad daje tačne odgovore onima koji su izvan sobe, bez ikakvog razumevanja značenja napisanog (Searle 1980). Pomenuti eksperiment vodi do zaključka da ispravni odgovori superračunara/ „jake AI" nisu isto što i ljudsko razumevanje problema, čime se osporava pretpostavka o mašinama kao svesnim entitetima (Searle 1980).

Kopča između Ostinovih teza i Serlove problematike kineske sobe leži u nemogućnosti da se dokaže unutarnje stanje drugog entiteta – bilo da je to čovek (Ostin) ili mašina (Serl). Otuda je ključna relacija između Ostinove tematike tuđih svesti i Serlove argumentacije iznesene povodom tzv. kineske sobe ona koja se zasniva na sumnji u zaključke o nekakvom unutarnjem stanju (uma), a koji su izvedeni na osnovu spoljašnjih manifestacija. Dok je Ostin skeptičan u pogledu ljudskog kapaciteta da sazna tuđu svest kroz ponašanje i govor, Serl pak kritikuje ideju da algoritamsko procesuiranje informacija može biti svesno. Obojica time podvlače jaz između onoga što je spolja vidljivo (ponašanje, govor ili algoritamska obrada) i onoga što se uistinu dešava iznutra (svest, razumevanje).

Povezivanje Ostinovih i Serlovih istraživanja sa idejom digitalnog totalitarizma može se posmatrati u kontekstu ispitivanja kako se simbolička mani-

4 Već na početku teksta „Minds, Brains, and Programs" („Umovi, mozgovi i programi") Serl pravi razliku između „slabe" (*weak*) i „jake" (*strong*) AI; „slaba" ili „oprezna" (*cautious*) veštačka inteligencija je ona koja predstavlja alat (u proučavanju uma), dok tzv. „jaka" nije samo puko sredstvo za izučavanje uma, već um sam. S tim u vezi, Serl se u tekstu bavi veštačkom inteligencijom koja razume ono što čini, a uz to poseduje i druge kognitivne dispozicije (Searle 1980).

pulacija potencijalno i realno koristi u modernim digitalnim sistemima za kontrolu i nadzor, bez bilo kakvog dubljeg razumevanja ili empatije prema ljudskim bićima i njihovim potrebama. Digitalni totalitarizam, kao sistem u kojem vladajuće strukture koriste napredne tehnologije, poput veštačke inteligencije, za praćenje, predviđanje i kontrolu ponašanja građana/ki, može biti izgrađen na sličnom principu kao „kineska soba". To podrazumeva da AI sistemi prate, analiziraju podatke i manipulišu njima bez stvarnog razumevanja njihovog značenja ili (društvenih, etičkih i drugih) posledica delovanja. U digitalnom totalitarizmu, dakle, mašine mogu donositi odluke koje neposredno utiču na život ljudi; te i takve odluke, međutim, nisu i ne mogu biti rezultat stvarnog promišljanja, već puke obrade podataka; inteligentne mašine ne mogu istinski moralno da prosuđuju, niti imaju svest o ljudskim autentičnim potrebama, pravima i slobodama. Kao i u Serlovom eksperimentu, AI sistemi „deluju" kao da razumeju ono što obrađuju, ali njihova manipulacija simbolima (u ovom slučaju informacijama o ljudima) može biti duboko dehumanizujuća.

Ono što je opasno u ovakvom sistemu jeste potencijal za zloupotrebu manipulacije podacima kojima raspolažu ili mogu raspolagati mašine. Ako elite na pozicijama moći koriste veštačku inteligenciju na ovaj način, one tada mogu opravdati odluke i akcije realizovane na osnovu „objektivnih" podataka, dok u stvari postoji ogroman prostor za greške, manipulaciju i etičke, političke i druge zloupotrebe ovakvih sistema – građani, odnosno žrtve, mogu biti etiketirani, sankcionisani ili kontrolisani na osnovu algoritama koji ne razumeju stvarni kontekst njihovih postupaka ili namera, čime se stvara sistem koji je mehanički i bezosećajan, a prikazuje se kao racionalan i pravičan. Uz to, ukoliko je kontekst kapitalistički (misli se prevashodno na ekonomiju, ne na ideološki predznak), on je i parazitski u odnosu na one koje nadzire, kontroliše i eksploatiše.

Šošana Zubof (Zuboff) u knjizi „Doba nadzornog kapitalizma" (*The Age of Surveillance Capitalism*), pozivajući se na Marksa (Marx) i prethodna vremena u kojima se kapitalizam, nalik na vampirske egzistencije, „hranio radom" proletera ostvarujući na taj način višak vrednosti u aktuelnom, tzv. nadzornom kapitalizmu, koji je, kako tvrdi autorka u potpoglavlju prvog dela pod nazivom „Šta je nadzorni kapitalizam?" (*What Is Surveillance Capitalism?*), „parazitski" i „autoreferentan" (Zuboff 2019), eksploatacija se nastavlja, ali ovoga puta u digitalnom prostoru. U njemu se, naravno, ostvaruje na daleko suptilniji način – upotrebom nemislećih mašina, koje osmatraju, usmeravaju i, najzad, eksploatišu naša iskustva. Nadzorni kapitalizam određuje

kapitalizam na taj način što se kontrola više ne ostvaruje nad sredstvima za proizvodnju, već nad alatkama za modifikovanje ljudskog ponašanja. Nosilac ovih modifikacija je, sugeriše Zubofova, kompanija *Google*, u čemu je prate i druge srodne internet-korporacije. Rečju, kapitalizam kao dominantna društveno-ekonomska formacija našeg vremena, posredstvom sveopšte digitalizacije koju prevashodno diktira ekonomija, predstavlja osnovnu pretnju dosad osvojenim pravima i slobodama čoveka; i mi dodajemo: tehnološka je osnova za uvođenje digitalizovanih oblika diktatura savremenog doba. Prema mišljenju Šošane Zubof, cilj nadziruće kapitalističkog poretka jeste u predviđanju i modifikovanju ljudskog ponašanja (*behavior modification*), kako bi se posredstvom kontrole ostvario profit.

Ali, moć nadziruće kapitalizma nije samo u poznavanju našeg ponašanja već i u monopolisanju informacija koje su posledica korisničkog ponašanja (lično ljudsko iskustvo); otuda se prikupljanje, kondenzovanje, analiza i dizajniranje ovih podataka tranformišu u tzv. bihevioralni višak, kojim se trguje na tržištu. (Digital Literacy and AI 2022)

Šta, međutim, mi kao obični istraživači znamo o tehnologijama programiranim da upravljaju našim ponašanjem, i kako uopšte možemo dospeti do takvih saznanja? Ukoliko imamo samo ona znanja koja počivaju na teorijskim pretpostavkama, te ih dokazujemo ili opovrgavamo putem misaonih eksperimenata, na primer, dok se na drugoj strani događa stvarnost o kojoj ništa ili vrlo malo znamo u pogledu upravljanja novim AI tehnologijama – na koji način da probijemo ili uklonimo saznajnu barijeru o nadziranju, kontrolisanju i upravljanju našim ponašanjem, ako je to, između ostalog, i poslovna tajna vodećih *hi-tech* kompanija? Na tragu ovih nedoumica i nepoznanica, pokušaćemo u nastavku teksta da povežemo Serlov eksperiment s konceptom „crne kutije" (*black box*), preuzetim iz Vinerove (Wiener) kibernetike.

Naime, u okviru kibernetike, „crna kutija" je konceptualizovana kao takav sistem za koji nije nužno da se njegova unutrašnja struktura poznaje, a da se on ipak može proučavati na osnovu ulaznih, kao i izlaznih informacija. Ilustracije radi, navešćemo ovde objašnjenje koje je dato na internet stranici Računarskog fakulteta u Beogradu. „Pri spominjanju termina 'crna kutija', mnogi pomisle na uređaje koji snimaju šta se dešava u avionima i izuzetno su važni za analiziranje događaja koji su se odigrali pre nesreće. U engleskom jeziku termin se koristi i za mala i minimalno opremljena pozorišta. Međutim, crna kutija je, takođe, važan pojam u svetu veštačke inteligencije. U oblasti veštačke inteligencije, crne kutije se odnose na interne procese koji

se odvijaju u sistemima veštačke inteligencije, a koji su potpuno nedostupni i nepoznati korisnicima. Kao što nam je poznato, takvim sistemima možete da date neki upit i dobijete neki odgovor, ali nemate uvid u sistemski kôd ili logiku koja je kreirala odgovor, odnosno, izlaz" (Računarski fakultet 2024).

Ako uporedimo Serlov eksperiment sa ukratko izloženim konceptom crne kutije, uočićemo sledeće podudarnosti. Serl u svom eksperimentu pokušava da dokaže da razumevanje jezika ili „svesti" nije samo stvar pravilne obrade simbola, već ono zahteva shvatanje značenja (semantike) rečenog ili napisanog. Eksperiment kineske sobe je demonstracija ideje da određeni program može simulirati znanje koje stvarno ne poseduje. Teorija „crne kutije", danas inače opštepoznata u inženjeringu, odnosi se na sisteme koji funkcionišu na takav način da ih korisnik ne razume u potpunosti. U crnoj kutiji poznate su ulazne i izlazne informacije, ali unutarnji procesi ostaju skriveni ili nepoznati. Ova teorija se često koristi za opisivanje veštačke inteligencije, pri čemu je izlaz (odluka ili rezultat) poznat, ali način na koji je veštačka inteligencija došla do odgovarajućeg ishoda neretko ostaje neproziran i teško razumljiv, čak i programerima i trenerima AI. Slično kao što u „kineskoj sobi" čovek unutar prostorije samo sledi pravila bez razumevanja onoga što radi, tako i „crna kutija" može da izvede složene procese bez mogućnosti da se otkrije kako i zašto su te odluke donete. U oba slučaja javlja se problem transparentnosti[5] i pitanje – kako možemo tvrditi da sistem (ljudski ili veštački) zaista razume ono što radi ako nam je pristup unutarnjim procesima ograničen ili potpuno nedostupan? Dakle, ova problematika može pokrenuti diskusiju o tome kako kompleksni sistemi, poput AI, mogu da simuliraju inteligenciju ili razumevanje, uz epistemološke granice nalik na one koje Serl naglašava u svom eksperimentu.

Jedan od pokušaja da se ovaj problem prevaziđe i korisnicima omogući uvid u to kako veštačka inteligencija koju koriste funkcioniše definiše se kao tendencija prelaska sa upotrebe sistema „crnih kutija" na algoritme koji su transparentni, tj. na sisteme tzv. belih (White box AI) ili staklenih (*Glass Box AI*) kutija. Ovaj prelaz treba da ostvari tzv. objašnjiva veštačka inteligencija (*explainable AI-XAI*), s ciljem da se osigura transparentnost i odgovornost delovanja putem razvoja onih modela koji mogu da objasne kako sistemi ve-

5 Prilikom korišćenja različitih modela mašinskog učenja AI, bilo koja od komponenti sistema (mašinskog učenja) može da bude sakrivena, odnosno da se „nalazi" u crnoj kutiji (nekad su to saznajni, a pokatkad i bezbednosni razlozi). Nasuprot ovome, postoje i sistemi relativno transparentni korisnicima i koje oni mogu da razumeju. „Oblast posvećena veštačkoj inteligenciji koja se može objasniti bavi se razvojem algoritama koje ljudi mogu bolje da razumeju, iako se to ne mora nazvati staklenom kutijom." (Računarski fakultet 2024)

štačke inteligencije donose vlastite odluke (Johal 2023). Tako se s XAI-jem može steći uvid u razloge koji stoje iza preporuka i odluka koje donosi veštačka inteligencija, čineći ih korisnicima verodostojnijim i razumljivijim. Pritom valja naglasiti da XAI ne treba da predstavlja samo zadovoljenje naše radoznalosti već i da omogući korišćenje koje je, simplifikovano rečeno – bezopasno. A posebno je važno da upotreba XAI osigura da veštačka inteligencija bude usklađena s ljudskim vrednostima i etikom. To bi bila, vrednosno gledano, jedna humanistički centrirana veštačka inteligencija. (Johal 2023)

U novijim člancima o mogućoj humanističkoj upotrebi objašnjive veštačke inteligencije (XAI) primećuje se promena u pristupu temi, koja se kreće od tehnocentričnog ka humanističkom i sociotehničkom fokusu razmišljanja. Tradicionalni XAI bio je, zapravo, usmeren na algoritamsku transparentnost, odnosno na objašnjavanje načina na koji AI donosi odluke otvaranjem „crne kutije". Ali razvoj složenijih modela s vremenom je pokazao da je teško „otvoriti" te sisteme, te u potpunosti razumeti njihove unutrašnje procese. Otuda je osnovno pitanje kada je reč o objašnjivosti AI – za koga ona jeste i treba da bude razumljiva? „Zapravo, izazovi dizajniranja i evaluacije AI-ja u 'crnoj kutiji' zavise od toga 'ko' je čovek u petlji." (Ehsan, Riedl 2020: 450) Rečju, postavlja se pitanje socijalne prepoznatljivosti rada AI, ali i mogućih zloupotreba.

Postavljeni izazov omogućio je pojavu tzv. ljudski centrirane objašnjavalačke veštačke inteligencije (*Human-Centered Explainable AI-HCXAI*), koja se fokusira na ljudski centrirane aspekte objašnjivosti, pozivajući se na argument da AI sistemi ne funkcionišu u vakuumu, nego u ljudskom okruženju (Ehsan, Riedl 2020). Utoliko su, kako se veruje, inteligentni sistemi deo „sociotehničkih" (*sociotechnical*) sklopova u kojima su ljudi ključni za interpretaciju i upotrebu tih sistema (Ehsan, Riedl 2020). Ovi sklopovi su zajednički za ljude i mašine, te se navodno razvijaju na obostranu korist. Stoga momenat objašnjivosti zavisi ne samo od tehnološke transparentnosti rada sistema nego i od načina na koji se veštačka inteligencija koristi i tumači u društvenom kontekstu. Drugim rečima, upotrebom HCXAI zagovara se pristup koji u sebe uključuje vrednosti korisnika i dizajnera,[6] čime se predočava da objašnjenja ne moraju dolaziti samo iz algoritama već se mogu naći i izvan njih; jer tu se nalaze ljudi i humano okruženje ključni za tumačenje i donošenje odluka u stvarnom svetu (Ehsan, Riedl 2020).

6 Ovaj pristup nagoveštavaju dve strategije koje približavaju dizajnere i korisnike AI, a to su participativni dizajn (*participatory design, PD*) i dizajn senzitivan na humane vrednosti (*value-sensitive design, VSD*). (Ehsan, Riedl 2020: 462)

Iz rečenog se vidi da HXAI načelno teži cilju da veštačka inteligencija bude transparentnija i prilagođenija čoveku, zahvaljujući ljudskim objašnjenjima odluka koje AI donosi; ali ako iza te transparentnosti stoje nedobronamerni, manipulativni procesi i interesi, onda se može postaviti pitanje integriteta i HXAI i osoba, odnosno kompanija i režima koji iza nje stoje. Kod ovakvih sistema postoji mogućnost zloupotrebe, tj. da neko pokuša da iskoristi objašnjenja AI sistema za manipulaciju, što predstavlja, naravno, ozbiljan problem jer se informacije mogu (zlo)upotrebiti kako bi se suptilno uticalo na ljude. U tom kontekstu sagledano, postavlja se pitanje da li neko zaista želi da pomogne da korisnik razume procese donošenja odluka, ili pak nastoji da oblikuje njegovo razmišljanje i, konsekventno, ponašanje u smeru koji mu odgovara? Ova prepostavka vraća raspravu na ideju o mogućoj zloupotrebi neznanja i informacija, pošto čak i tehnologija koja izgleda kao da je u službi transparentnosti može imati skrivenu agendu delovanja. Na osnovu predočenog, relevantno pitanje jeste šta je prava svrha tih objašnjenja i ko ih uistinu kontroliše?

Pretpostavimo da je manipulacija sistemima AI profitabilna. Tada je teško zamisliti da bi onaj ko profitira od veštačke inteligencije bio iskren u pogledu upotrebe određenih metoda. Upravo zato što je deo tog sistema, manipulator (personalni, korporativni, institucionalni) u situaciji je da koristi objašnjenja kao fasadu, odnosno da uverava korisnike da je sve transparentno i u njihovu korist, dok se u stvari smera na kontrolu i manipulisanje. U takvom scenariju objašnjenja nisu stvarna objašnjenja, već strategija za stvaranje lažnog osećaja sigurnosti ili poverenja u AI. Takvim postupcima kreira se iluzija razumevanja, kako bi se smanjila sumnja korisnika, u odsustvu stvarnog razotkrivanja suštinskih mehanizama manipulacije. Ovaj model može postati samoodrživ – sistem korisnicima daje informacije koje ograničavaju njihovu sposobnost da prepoznaju manipulaciju, dok u isto vreme simulira da im pomaže da bolje razumeju procese odlučivanja. Drugim rečima, takav sistem može funkcionisati kroz stvaranje iluzije izbora, dok korisnike drži u okviru unapred zadatih opcija „odlučivanja".

A ko čini takav sistem koji počiva na asimetriji znanja (i moći)? Uvešćemo, na ovom mestu, argumentaciju zasnovanu na podeli na tzv. algoritamske elite, s jedne, i korisnike neprozirnih AI sistema, s druge strane. Ovakva terminologija neretko se koristi u kritičkim analizama koje se bave neprozirnošću algoritamskih sistema, kao i moći zasnovanoj na njima, kao deo savremenog diskursa o tehnološkoj i političkoj moći, sagledanoj u kontekstu delovanja velikih tehnoloških kompanija, kao i onih režima koji kontrolišu i kreiraju

algoritme što oblikuju aktuelne društvene tokove. Poimanje tehnoloških (algoritamskih) elita referiše, kako samo ime kaže, na ograničen broj ljudi ili organizacija (najčešće velikih tehnoloških kompanija) koji kreiraju i kontrolišu algoritamske sisteme, odnosno poseduju znanje i kapacitete da projektuju algoritme koji upravljaju ključnim aspektima savremenog života: finansijskim tržištima, društvenim mrežama, pretragama na internetu, zapošljavanjem, kreditiranjem, pravosudnim procesima, obrazovanjem, medijima i dr. Ove elite, nasuprot tzv. običnim korisnicima, poseduju neproporcionalnu moć i ostvaruju veliki društveni uticaj i dobit koristeći algoritamske alate u cilju donošenja važnih odluka, i to na način koji nije transparentan široj javnosti. Njihova moć proizlazi iz pristupa ogromnim količinama podataka, te mogućnosti da ih koriste za manipulisanje tržištima, društvenim interakcijama i političkim procesima.

Nije redak slučaj da pomenute elite izbegavaju nadzor države ili javnosti, pošto su njihovi algoritmi, kako smo rekli, neprozirni ili zatvoreni („crne kutije"). Veoma mali broj ljudi razume kako oni zaista funkcionišu, što omogućava tehno-eliti kontrolu nad procesima koji utiču na svakodnevni život ogromnog broja ljudi. Ukratko, ove elite kontrolišu društvene infrastrukture na način koji običnim korisnicima onemogućava pristup informacijama o tome kako algoritmi dolaze do određenih zaključaka. To stvara značajnu prednost elitama u odnosu na korisnike i omogućava im da održavaju monopol nad algoritamskom infrastrukturom. Nadalje, algoritamske elite su koncentrisale moć u rukama tehnoloških giganata kao što su *Google*, *Facebook*, *Amazon* i sl. Ove kompanije kontrolišu infrastrukturu putem koje se širom sveta plasiraju informacije, proizvodi i usluge. Jasno je da algoritamske elite mogu da zloupotrebe moć za vlastitu korist, bez obaveze da odgovaraju pred zakonom ili poštuju pravila koja štite društvene interese.[7] Uz to, ovi algoritmi mogu postupati diskriminatorno, širiti dezinformacije ili targetirati određene grupe ljudi bez njihovog znanja ili pristanka.

U kontekstu analize o digitalnom totalitarizmu, valja kritički reflektovati o algoritamskim elitama, pošto se moć algoritama da utiču na ponašanje ljudi i političke procese može smatrati oblikom tehnološkog autoritarizma u kojem oni preuzimaju ulogu odlučivanja, ranije rezervisanu za demokratske institucije. Neodgovornost ovih elita i nemogućnost da obični građani nadziru te

7 Poznat je, recimo, slučaj *Fejsbuka* (Meta) povodom zloupotrebe i masovnog prodavanja podataka korisnika, uticaja na izbore, širenja lažnih vesti, cenzure, nasilja, zloupotrebe dece (rasprava u američkom Kongresu) i izbegavanja plaćanja poreza, za koje ni izvršni direktor *Fejsbuka* Mark Zakerberg (Zuckerberg), a ni njegova kompanija nisu odgovarali.

procese direktno dovodi do pojave digitalnog totalitarizma. Ovakva situacija pretpostavka je realizovanja tzv. društva crne kutije, o čemu je kritički pisao Frank Paskvali (Pasquale) u istoimenoj knjizi „Društvo crne kutije" (*The Black Box Society*) (Pasquale 2015).

Paskvali u pomenutoj knjizi detaljno analizira kako neprozirni algoritmi kontrolišu ključne društvene procese, prevashodno finansijske tokove, informacije i pravosuđe. Autor se u studiji o crnim kutijama podrobno bavi istraživanjem strategija koje omogućavaju da se one drže zatvorenim za javnost. U tom smislu, imajući u vidu da je znanje moć,[8] Paskvali tvrdi da dekonstrukcija crnih kutija velikih podataka (*Big Data*) nije jednostavan poduhvat (Pasquale 2015: 6). Čak i ako bi algoritamska elita bila spremna da izloži svoje metode javnosti, savremeni internet i bankarski sektor predstavljali bi i dalje prepreke za poznavanje metoda manipulacije u odnosu na krajnje korisnike (Pasquale 2015: 6). Jer zaključci do kojih internet korporacije i bankari dolaze, recimo o produktivnosti zaposlenih, relevantnosti veb stranica ili povoljnim ulaganjima, sintetišu se putem složenih formula koje su „osmislile legije inženjera i koje čuva falanga advokata" (Pasquale 2015: 6).

Razumevanje i praćenje razvoja AI tehnologija s obzirom na fenomen crnih kutija važno je kako bi se osiguralo da se one koriste na način koji podrazumeva poštovanje ljudskih prava i demokratskih vrednosti. Međutim, ako se digitalne tehnologije, a pre svega veštačka inteligencija zloupotrebe kako bi se ostvarila potpuna kontrola nad informacijama, komunikacijom i ponašanjem ljudi, u tom kontekstu posmatrano društvo crnih kutija bi moglo da anticipira ili predstavlja ranu fazu implementacije digitalnog totalitarizma. To bi, ujedno, značilo da tehnologije poput AI, koje funkcionišu kao crne kutije, omogućavaju korporacijama ili vlastima da donose netransparentne odluke bez preuzimanja odgovornosti, te da ograniče ljudske slobode i zloupotrebe moć (znanja) na štetu većine ljudi.

Preciznije rečeno, percepcija društva crnih kutija u kontekstu umreženih digitalnih tehnologija, posebno AI, može se posmatrati kao novi oblik mrežnog („rizomskog") digitalnog poretka u kojem su ljudi, institucije i sistemi povezani složenim i često nevidljivim mrežama odlučivanja. U takvom društvu, umrežavanje crnih kutija predstavljalo bi višestruke slojeve interakcija i komunikacija, koje su neprozirne i teško razumljive spoljašnjim posmatračima.

8 Prizivajući Bekona (Bacon), Paskvali nastavlja da elaborira stav o odnosu znanja i moći kroz prizmu delovanja neprozirnih algoritama, tvrdeći da „ispitivati druge, a izbegavati ispitivanje sebe jeste jedan od najvažnijih oblika moći." (Pasquale 2015: 3)

U tom slučaju, unapređena veštačka inteligencija postala bi jezgro upravljanja umreženim društvom, ali tako da način na koji ona donosi odluke često nije transparentan.

Nadalje, crne kutije bi, u ovom interpretativnom kontekstu, bile ne samo algoritmi već bi ih činile i sistemske odluke koje veštačka inteligencija donosi, kao i način na koji se te odluke primenjuju. To znači da bi se svakodnevne aktivnosti pojedinaca mogle kontrolisati i moderirati putem AI sistema, bez jasnog uvida u kriterijume ili procese koji oblikuju te odluke. Pojedinci unutar sistema ne bi mogli da razumeju te procese ili upravljaju njima, čime bi društvo ušlo u fazu u kojoj algoritmi, umesto ljudi, postaju autoriteti znanja i društvene prakse. Ovakva dinamika vodila bi, ako se ne reflektuje i ne kontroliše, digitalnom totalitarizmu u kojem su ljudi podložni nadzoru, te kontrolisani kroz sistem koji ne razumeju, dok istovremeno ne znaju kako da se zaštite od njegovih efekata.

No, pitanje transparentnosti funkcionisanja crnih kutija je samo površinsko, dok je pravi problem u tehno-kapitalizmu, koji koristi tehnologiju kao alat za kontrolu, eksploataciju i održavanje sistemske moći kapitalističkog društveno-ekonomskog poretka. Drugim rečima, digitalna mreža crnih kutija ne mora da bude transparentna; ali ona svakako funkcioniše na način koji omogućava centralizovanu moć i profit u rukama tehnokratskih i kapitalističkih elita. Ovaj novi, lucidniji način nadziranja i kontrolisanja pojedinaca i čitavih zajednica uključuje sofisticirane oblike manipulacije, pri čemu crne kutije (AI sistemi, algoritmi) funkcionišu u službi korporativnog i kapitalističkog interesa, držeći korisnike u potlačenoj poziciji neznanja i robovanja. Tako tehnologija postaje sredstvo za dalju eksploataciju resursa. Umesto tradicionalnog rada, kapitalizam sada koristi digitalne sisteme za ekstrakciju vrednosti iz podataka, algoritama i „pametnih" digitalnih interakcija. Umrežavanje crnih kutija unutar kapitalističkog sistema služi za efikasnije upravljanje i kontrolu radne snage i potrošača. Ovaj se nadzor ne vrši samo putem tradicionalnih mehanizama kontrole već kroz tehnologiju koja poseduje autonomne moći odlučivanja. Time se stvara iluzija slobode za pojedince, dok su oni, u stvari, sve više integrisani u mreže nadzora, kontrole i eksploatacije.

Stoga verujemo da je demistifikacija pomenutih procesa, kao i njihovo kritičko promišljanje, prvi korak u suočavanju s problemima koje donosi tehno-kapitalizam.[9] U tom kontekstu, razumevanje kako se tehnologija veštačke

9 Na jednom drugom mestu govorili smo, a povodom savremenih medija, o „tehnološki podržanom ludilu kapitalizma". (Vuksanović 2022: 3292)

inteligencije koristi za uspostavljanje i održavanje moći, krećući se ka porobljavanju i digitalnom totalitarizmu, ali i mogućnosti za drukčije i humanije pristupe ovoj problematici, može pomoći u izgradnji pravednijeg i demokratičnijeg sistema koji teži ostvarenju većeg stepena slobode, kako za pojedinca tako i za društvene zajednice našeg vremena.

Odabrana bibliografija s netografijom

- Austin J. L. (1980). Tuđe svesti. U: *Svest i saznanje*, Beograd: Nolit, 141-184.
- Defining Surveillance Capitalism (2022). In: *Digital Literacy and AI*, Pepperdine Libraries, na stranici: https://infoguides.pepperdine.edu/digitalliteracy/surveillancecapitalism, pristupljeno: 15. 10. 2024.
- Deset godina kasnije: Edvard Snouden i njegova otkrića (2023). *DW*, na stranici: https://www.dw.com/sr/deset-godina-kasnije-edvard-snouden-i-njegova-otkri%C4%87a/a-65837392, pristupljeno: 4. 10. 2024.
- Ehsan, U. Riedl, M. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach, na stranici: https://www.researchgate.net/publication/340438092_Human-centered_Explainable_AI_Towards_a_Reflective_Sociotechnical_Approach, pdf.
- Godfrey-Smith, P. (2017). *Other Minds: The Octopus and the Evolution of Intelligent Life*. London: William Collins.
- Johal, A. (2023). From Black Box to Glass Box: The Evolution of Explainable AI: The Human Side of AI: Why Explainable AI is Essential for Human-Centered AI, na stranici: https://medium.com/geekculture/from-black-box-to-glass-box-the-evolution-of-explainable-ai-c4b-932c9fe94,pristupljeno: 16. 10. 2024.
- Harris, J. (2019). Reading the minds of those who never lived. Enhanced beings: The social and ethical challenges posed by super intelligent AI and reasonably intelligent humans., *Cambridge Quarterly of Healthcare Ethics*, *28*(4), 585–591.
- Nyholm, S. (2019). Other minds, other intelligences: the problem of attributing agency to machines. *Cambridge Quarterly of Healthcare Ethics*, *28*(4), 592–598; and pdf, 1–8.
- Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, Massachusetts, London, England: Harvard University Press.

- Spannos, Ch. (2016). Mass Surveillance and *Smart Totalitarianism*. *ROAR*, Issue #4: 'State of Control', na stranici: https://roarmag.org/magazine/mass-surveillance-smart-totalitarianism/, pristupljeno: 4. 10. 2024.
- Šta je veštačka inteligencija u crnoj kutiji? (2024). Računarski fakultet Unverziteta Union u Beogradu, na stranici: https://raf.edu.rs/citaliste/sta-je-vestacka-inteligencija-u-crnoj-kutiji/, pristupljeno: 16. 10. 2024.
- Vuksanović, D. (2022). Philosophy in the Time of Media and Technological-information Madness. *In medias res: časopis filozofije medija*, *11*(20), Zagreb: Centar za filozofiju medija i mediološka istraživanja, str. 3292., na stranici: https://www.centar-fm.org/inmediasres_eng/index.php/divna-vuksanovic-philosophy-in-the-time-of-media-and-technological-information-madness, pristupljeno: 20. 10. 2024.
- Zuboff, Shoshana (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, (uvod), we.riseup.net Crabgrass, na stranici: https://we.riseup.net, Pdf.

Divna Vuksanović
Faculty of Dramatic Arts in Belgrade

ARTIFICIAL INTELLIGENCE AND DIGITAL TOTALITARIANISM

Abstract

The widespread digitalization in the technologically advanced world has been undertaken without sufficient prior planning or reflection. As a venture with totalizing ambitions, compulsive digitization promises a technological paradise: the Internet of Things, the reign of transhumanist values, and the rise of a robotics and artificial intelligence empire. Critics of these processes foresee the potential for new forms of totalitarianism, or modern-day enslavement, realized through the comprehensive deployment of artificial intelligence within the digitization phenomenon. This paper seeks to examine the (in)capabilities of intelligent machines and, more precisely, to explore how AI philosophy responds to these ongoing societal developments.

Keywords

digitization, artificial intelligence, totalitarianism, knowledge, power, capitalism

Primljeno: 21. 10. 2024.
Prihvaćeno: 19. 5. 2025.