DOI
https://doi.org/10.18690/um.pf.4.2023.1

ISBN
978-961-286-774-4

1 ETIKA AVTOMATIZACIJE, DIGITALIZACIJE IN UMETNE INTELIGENCE: OBSTOJEČE DILEME, TVEGANJA IN REŠITVE

NIKO ŠETAR
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
niko.setar1@um.si

V spodnjem prispevku preučujemo etične in moralne vidike obstoječih in nastajajočih tehnologij. Pri tem začenjamo z analizo avtomatizacije delovnih mest in implikacij avtomatizacije za človeštvo v preteklosti, sedanjosti in prihodnosti. Nadaljujemo z obravnavo pojava naraščajoče digitalizacije, s težavami, s katerimi se v njenem kontekstu soočajo tako ustvarjalci kot uporabniki, s poudarkom na spletni anonimnosti, človekovih pravicah in algoritemski diskriminaciji, ter možnih etičnih in pravnih rešitvah za nastale in nastajajoče težave v praksi. Ker sodobna digitalizacija pogosto uporablja umetno učenje in preprosto umetno inteligenco, se prispevek v nadaljevanju osredotoča na etično obravnavo avtonomnih sistemov, kot so samovozeča vozila in samodejna orožja, npr. 'droni'. Iz problematike, ki jo obravnavamo v tem sklopu, se navežemo na izzive, ki jih predstavlja razvoj višje, človeku podobne umetne inteligence, vključno z bolj futurističnimi scenariji, kot so singularna superinteligenca, moralni status človeku podobne umetne inteligence kot osebe, ter verjetnost nastanka tovrstne inteligence, možne etične rešitve ter dileme, ki ostajajo nerešene.

Ključne besede:
avtomatizacija, digitalizacija, umetna inteligenca, avtonomni sistemi, anonimnost, diskriminacija

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/
10.18690/um.pf.4.2023.1

ISBN
978-961-286-774-4

1 The Ethics of Automatization, Digitalization and Artificial Intelligence: Existing Dilemmas, Risks and Solutions

Niko Šetar
University of Maribor, Faculty of Arts, Maribor, Slovenia
niko.setar1@um.si

Keywords:
automatisation,
digitalisation,
artificial intelligence,
autonomous systems,
anonimity,
bias

In the following article, we look into ethical and moral aspects of existing and emerging technologies. In doing so, we begin by analysing automatisation of workplaces and implications of such automatisation for humanity in the past, present, and future. We continue with the phenomenon of increasing digitalisation, and what kind of issues its creators and users are facing, emphasising on web anonimity, human rights, and algorithmic bias, as well as possible ethical and legal solutions for existing and emerging problems. Seen as contemporary digitalisation often uses machine learning and rudimentary artificial intelligence, our article continues in that light, with ethical consideration of autonomous systems, such as self-driving vehicles and autonomous weapons, e.g. drones. From the problematic that we address in this section, we relate to the challenges presented by the development of higher, human-like artificial intelligence, including futuristic scenarios like singular superintelligence, moral status of human-like artificial intelligence as a person, considering the probability of such intelligence's existence, possible ethical solutions, and persisting dilemmas.

University of Maribor Press

1.1 Uvod

Človeštvo je do točke razvoja, na kateri se nahajamo danes, napredovalo zaradi svoje inovativne narave, deljenja znanj in informacij, pogosto pa tudi zaradi čiste sreče, in na račun marsikaterega spodletelega eksperimenta. Vedno, kadar se je pojavila nova tehnologija, so se z njo pojavili tudi skeptiki, paranoiki, in ostali, ki so o njej verjeli eno ali drugo neresnico: v poznem 19. stoletju je veljal Teslin dvofazni električni tok, ki ga danes uporablja cel svet, za smrtno nevarnega, pretežno zaradi laži, ki jih je širil Thomas Edison, da bi uveljavil in profitiral od svojega enosmernega toka; prav te dni na socialnih omrežjih opažamo širjenje nenavadnih teorij, kot tudi teorij zarote, o (stranskih) učinkih 5G omrežja, ki nimajo otipljive znanstvene podlage. To pa seveda ne pomeni, da nove tehnologije niso brez tveganj: za posledicami radioaktivnega sevanja so umrli mnogi raziskovalci, vključno s pionirko Marie Sklodowsko Curie, Einstein in Oppenheimer pa sta opozarjala proti uporabi jedrske tehnologije v vojaške namene – Oppenheimer sicer šele po tem, ko je na lastne oči videl uničenje tesntega poligona po detonaciji eksperimentalne bombe Trinity.

Pri vsem tem je jasno, da so pri novih dognanjih, ki jih morda še ne razumemo v celoti, potrebni določeni preventivni ukrepi; pri tistih, ki jih razumemo, in smo uspeli identificirati napake, tveganja, in podobno, pa je nujna vpeljava ukrepov, ki bodo ta tveganja omejili oziroma odpravili. Pri tem lahko gre za mehanske napake, ki jih je moč odpraviti s tehničnimi rešitvami, ali pa drugovrstna tveganja, na primer za okolje, družbo, ali posameznika, ki upravlja ali je v stiku z novo tehnologijo.

1.2 Tehnološki razvoj in povezana tveganja

1.2.1 Regulacija razvoja in novih tehnologij: etični vidik

Brey trdi, da je glavni problem obravnavanja in regulacije novih tehnologij negotovost, tj. negotovost o lastnostih in možnih posledicah še nerazvite tehnologije[1]. Pri odgovornem obravnavanju etičnega problema, ki je negotov, se je potrebno izogniti tako pretirani špekulaciji, kot tudi ideji, da zaradi negotovosti dotičnega problema ne moremo obravnavati, saj lahko to vodi bodisi v prenizko reguliranost razvoja, ali pa v pretirane regulacije, ki lahko aktivno zavirajo napredek.

[1] Brey, P.: Anticipatory Ethics for Emerging Technologies, v: Nanoethics, 6 (2012) 1, str. 1-13.

Dva pristopa, ki se v splošnem uporabljata za tovrstno etično obravnavo, sta generični pristop in napovedni pristop; generični pristop obravnava splošne oz. evidentne lastnosti nove tehnologije (npr. problem radioaktivnosti pri jedrski energiji), napovedni pristop pa poskuša predvidevati prihodnjo rabo in možne posledice nove tehnologije.[2]

Brey v svojem članku kritizira pretekle metode etične obravnave novih tehnologij, vključno z najbolj priznano metodo ETICA, ki temelji na obravnavi problema na podlagi združevanja različnih napovedi o isti zadevi pod domnevo, da je vsaka posamezna napoved pristranska, zaradi česar je potrebno vzeti v poštev maksimalno število možnih napovedi in etičnih obravnav. Metodi ETICA Brey očita, med drugim, da večina njenih analitičnih zmožnosti temelji zgolj na generičnih lastnostih tehnologije, in da je nezadostna za temeljito obravnavo.[3]

Brey predlaga metodo ATE (Anticipatory Technology Ethics), ki vključuje model analize novih tehnologij na treh nivojih: na nivoju tehnologije (generične lastnosti), artifakta (specifični predmeti, sistemi in postopki, ki uporabljajo tehnologijo), in aplikacije (različne rabe artifakta). Na vsakem nivoju je potrebno izpeljati dvostopenjsko etično analizo, pri čemer stopnji povzema po metodi ETICA. Prva stopnja, identifikacija, je namenjena identifikaciji etičnih problemov v skladu z etičnim seznamom, ki zaobjema štiri sekcije. Prva je »Škoda in tveganja«, ki nadalje vključuje telesne poškodbe, psihološke učinke, okoljsko škodo, itd.; druga »Pravice«, ki med drugim obsega različne svoboščine, dostojanstvo, avtonomijo in zasebnost; tretja »Sodstvo«, in četrta »Dobrobit in skupno dobro«. Sledi druga stopnja analize, evalvacija, kjer se oceni resnost negativnih vplivov na elemente, vsebovane na etičnem seznamu, in izvedba možnih ukrepov.[4]

Na pomanjkljivosti teorije ETICA se sklicuje tudi Nathan,[5] ki pa se svoje metode reševanja problemov tehnološke in inovacijske etike loti tako, da definira štiri glavne interesne skupine, na katere vpliva nova tehnologija: to so direktorji, stranke, mediji in vlada. Vsaka izmed naštetih skupin ima svoj tip (dominantna, odvisna, zahtevna

[2] Ibidem.
[3] Ibidem.
[4] Ibidem.
[5] Nathan, G.: Innovation process and ethics in technology: An approach to ethical (responsible) innovation governance, v: Journal on Chain and Network Science, 15 (2015) 2, str. 119-134.

in dominantna, kot si sledijo zgoraj), interese, pravice, odgovornosti in etična tveganja.

V starejšem članku se Di Norcia pri vzpostavitvi sheme razvojnih težav in rešitev opira na tako imenovani Tehnološki Ciklus, ki poteka v šestih stopnjah: inovativni preboj, razvoj in variacija, širjenje in izbor, masovna raba/standardizacija, zrelost/prevlada in zaton/nov preboj.[6] Temu sledi vzporeden Ciklus težav. Ker nas v tem prispevku zanimata preboj in razvoj, se pravi prvi dve stopnji Tehnološkega ciklusa, bomo omenili le težavi, ki jima pritičeta. Pri inovativnem preboju je glavna težava pomanjkanje poročil o možnih napakah, težavah in etičnih dilemah, pri razvoju in variaciji pa pride, po Di Norcii, šele do začetka razumevanja, kjer pa še zmeraj prevladujeta pristranskost in zanikanje težav s strani tistih, ki sodelujejo pri razvoju. Večino rešitev, ki jih predlaga Di Norcia, so novejše metodologije že nadgradile, smiselno pa je izpostaviti en predlog, ki ga večina drugih avtorjev, ki se osredotočajo bolj nad samo-nadzor podjetij, ki razvijajo nove tehnologije, zanemari: neodvisne organizacije, namenjene nadzoru etike inovacij in tehnološkega razvoja.

Čeprav se izhodiščne točke in metodologije opisanih pristopov razlikujejo, so skupne točke očitne. Vsi našteti viri, in viri, na katere se ti sklicujejo, navajajo enakopravno upoštevanje interesov vseh vpletenih interesnih skupin, kar se pri razvoju tehnologije v dobro kapitala (po Di Norcii skupine direktorjev), prej kot karkoli drugega, pogosto ignorira. Predvsem se pogosto zaobidejo parametri psihološke, okoljske in socialne škode kot posledice novih tehnologij (ali nezadostne regulacije) tudi takrat, ko se ekonomsko in fizično tveganje dosledno upoštevata.

1.2.2 Digitalizacija vsakdanjega življenja in njene pasti

V 21. stoletju je ena izmed mnogih nastajajočih tehnologij umetna inteligenca, ki se nahaja nekje na prevesu 1. in 2. stopnje Di Norcievega Tehnološkega ciklusa, pri čemer njeni teoretični nasledniki, kot so resnična, človeku-podobna ali singularna umetna inteligenca (v nadaljevanju UI), v ta ciklus še niso niti vstopili, njeni predniki pa so že napredovali po omenjeni lestvici: osnovni roboti in naučene naprave se nahajajo na 3.-4. stopnji, splošna digitalizacija pa že kar na 5. stopnji.[7] Kakšne so

[6] di Norcia, V.: Ethic, Technology Development, and Innovation, v: Business Ethics Quarterly, 4 (1994) 3, str. 235-252.
[7] Glej: di Norcia, Ethic, Technology Development and Innovation.

dileme, s katerimi se soočamo pri umetni inteligenci, oziroma se bomo soočali pri njenih višjih oblikah, in kaj nas lahko o njihovem preventivnem reševanju nauči naša dosedanja obravnava osnovnejše robotike in digitalizacije?

Capurro navanja, da sega dialog o digitalizaciji in njenih implikacijah nazaj v 80. leta prejšnjega stoletja, ko se v luči napredka računalniške tehnologije poraja ideja 'informacijske družbe' – ideja, ki jo je leta 1993 uresničila CERN-ova deklaracija, da bo Berners-Leejev »World Wide Web« odslej prosto dostopen.[8] Od takrat se je digitalna tehnologija, povezana med sabo s pomočjo globalnega interneta, razširila po vsem svetu, in postala del vsakdanjega življenja večine svetovnega prebivalstva. Integracija digitalnega sveta in moderne družbe je postala tako tesna, da je Floridi (2015) skoval izraza »onlife« in »offlife«, ki označujeta internetno in ne-internetno življenje vsakega posameznika.[9] Capurro trdi, da je ta pretirana integracija in pretirana povezanost sveta eden izmed dejavnikov, ki negativno vplivajo na človeško dostojanstvo iz mnogih razlogov, med drugim zato, ker internet omogoča primerjavo povprečnega uporabnika z nedosegljivim idealom. Nadalje opaža tudi, da mnoge digitalne storitve predstavljajo nevarnost za pravice posameznika; možnost neprestanega nadzora ogroža uporabnikovo zasebnost, tarčno oglaševanje pa njegovo avtonomijo odločitev.[10]

Royakkers postreže s številnimi skrb vzbujajočimi primeri nenadzorovane digitalizacije, nekaj od teh bomo na tem mestu povzeli.[11] Vdor v zasebnost ilustrira s primerom opozorila, najdenega v (46 strani dolgih) navodilih za uporabo televizije proizvajalca Samsung, ki pravi: »Prosimo, da se zavedate, da če vaše besede [izgovorjene v bližini televizorja] vsebujejo osebne ali druge občutljive podatke, bodo ti podatki med tistimi, ki bodo posneti in posredovani tretji osebi.«[12] Podobno avtor povezuje Googlova očala, izdana leta 2013, ki pa niso nikoli vstopila v širšo komercialno rabo, in nevarnost tako imenovanega »Little Brother« scenarija, v katerem sicer ne gre za nadzor vlade nad posamezniki, ampak za navzkrižni nadzor posameznikov in podjetij nad drugimi posamezniki in podjetji – vdor v Google očala

[8] Capurro, R.: Digitalization as an ethical challenge, v: AI & Society, 32 (2017), str. 277-283.
[9] Ibidem, str. 279.
[10] Povzeto po: Ibidem.
[11] Royakkers, L. in ostali: Societal and ethical issues of digitalization, v: Ethics and Information Technology, 20 (2018), str. 127-142.
[12] Ibidem, str. 129.

namreč omogoča dostop do celotnega uporabnikovega vidnega polja in slušnih zaznav.[13]

Kršitve osebne avtonomije se lahko pojavijo na več načinov; Royakkers med drugim ilustrira na videz banalen, a brez dvoma mogoč primer, ki ga imenuje tehnološki paternalizem. V njegovem primeru gre preprosto za pameten hladilnik, ki, ko v njem zmanjka sira, samovoljno naroči sir z nižjo vsebnostjo maščob, ker mu je druga naprava sporočila, da je uporabnikov holesterol previsok.[14]

Pojavi se tudi problem svobode izražanja, ki jo lahko kršijo razni algoritmi za cenzuro, lahko pa pride tudi do obratnega pojava, ki se v zadnjih letih očita socialnim omrežjem, in sicer, da do prevelike mere dopuščajo širjenje lažnih novic.[15] Izredno zaskrbljujoč je tudi problem splošne varnosti nekatelih digitalnih naprav – Royakkers navaja, da je Univerza v Teksasu leta 2012 demonstrirala, kako je z replikacijo digitalne identitete uporabnika (ang. Spoofing) mogoče precej enostavno vdreti v vojaški dron. Kljub temu, da je od te demonstracije do časa pisanja tega prispevka minilo že nekaj časa, in imajo današnji vojaški droni brez dvoma izboljšanje varnostne sisteme, verjetno obstajajo tudi izboljšane metode digitalnega vdora v tovrstne sisteme.[16]

1.2.3 Digitalizacija in umetna inteligenca

Problematika se le še zaostri, ko jo razširimo na domeno umetne inteligence. Kot opaža Anderson, se pojavijo težave že pri 'primitivnejši' umetni inteligenci, in ne šele pri človeku-podobnih ali superinteligentnih sistemih.[17] Avtorica se pri tem sklicuje na eksperiment, pri katerem je robot, namenjen pomoči starostnikom, zaradi neprimernega časa ali pretirane paternalizacije večkrat užalil varovanko. Ta robot sicer ni bil dovolj napreden, da bi lahko s slabo načrtovanim dejanjem povzročil škodo, a si je enostavno predstavljati, da bodo naprednejše, pametnejše umetne inteligence postavljene pred resnejše dileme in pomembne odločitve, kjer bi lahko

[13] Povzeto po: Google Glass and Privacy, Electroic Privacy Information Center, <epic.org/privacy/google/glass/#Privacy%20Interests> (29. 5. 2020).
[14] Povzeto po: Royakkers i.o., Societal and ethical issues.
[15] Povzeto po: Heijinen, I.: Fake News Social Media, EuropCom 2017 – Media Literacy Workshop.
[16] Povzeto po: Royakkers i.o., Societal and ethical issues.
[17] Anderson, S. L.: Machine Ethics, v: Anderson, J. M. in Anderson, S. L.: Machine Ethics, Cambridge University Press, Cambridge 2016, str. 1-19.

napaka v presoji vodila v katastrofo. Zato je pomembno, da so etična načela vključena v razvoj umetne inteligence. Anderson pri tem poudarja, da morajo v ta namen tehniki, ki razvijajo umetno inteligenco, prisluhniti izvedencem za etiko, saj je etika poglobljena disciplina, normativna etika pa se bistveno razlikuje od intuitivne etike povprečnega posameznika.

Kompleksnost etike je eden izmed glavnih problemov pri razvoju zanesljive umetne inteligence: medtem, ko je utilitarizem dejanja sprejemljiv kandidat za uporabo v umetni inteligenci zaradi svoje objektivnosti, mu je mogoče očitati, da odobrava žrtvovanje posameznika v dobro družbe; deontologija, ki tega nikoli ne bi dopustila, pa preveč zanemarja posledice dejanj. Teorija, ki bi ustrezno združevala načela zgornjih dveh, in bi bila del programa človeku-podobne, empatične umetne inteligence, bi bila najboljša rešitev.[18]

Naslednji problem, ki se pojavi, je ali lahko takšna umetna inteligenca sploh obstaja. Za etično ravnanje je potrebna intenca, predpogoja za katero pa sta zavestnost in svobodna volja. Mnogi predvidevajo tudi, da je za pravilno etično ravnanje potrebna tudi empatija, katere predpogoj je zmožnost čutenja in čustvovanja. Zaenkrat še ne vemo, ali lahko umetna inteligenca izpolni katerega izmed teh dveh pogojev.[19]

Bostrom in Yudkowsky predlagata nekaj pogojev, ki bi jih bilo potrebno izpolniti, da se čimbolj zmanjša tveganje, da nam umetna inteligenca uide izpod nadzora.[20] Prvi izmed teh je transparenca za pregled, ki zahteva, da imamo vpogled v notranje delovanje umetne inteligence, in da jo lahko ob kakršnemkoli nepravilnem delovanju pregledamo tako, da je mogoče odkriti vzrok nepravilnosti. Drugi pogoj je predvidljivost, ki sledi delno iz osnovnega programa umetne inteligence, delno pa po tem, da bi naj umetna inteligenca reševala probleme določenega tipa v skladu z replikacijo reševanja preteklih problemov tega tipa. Tretji pogoj je robustnost, ki naj bi onemogočala možnost digitalnega vdora v umetno inteligenco in vmešavanje v njeno delovanje.

[18] Povzeto po: Anderson, Machine Ethics.
[19] Ibidem.
[20] Bostrom, N. in Yudkowsky, E.: The Ethics of Artificial Intelligence, v: Ramsey, W. in Frankish, K.: Cambridge Handbook of Artificial Intelligence, Cambridge University Press, Cambridge 2011, str. 1-20.

Nadaljni problem je prva dva izmed teh pogojev združiti z idejo splošne umetne inteligence; ta dodatni pojem splošnosti pomeni, da umetna inteligenca ni usmerjena v opravljanje specifične naloge, ampak ima podoben, isti, ali celo širši razpon različnih nalog kot človek.[21] Hkrati je smisel umetne inteligence, da nadomesti človeka pri opravljanju naloge, torej da je pri tem od njega boljša – to pa je nemogoče doseči, če umetni inteligenci ne dovolimo samostojnega učenja, ampak jo omejimo na zmožnosti/znanje programerjev. Takšna umetna inteligenca, kot pravita Bostrom in Yudkowsky, nikoli ne bi premagala Kasparova v šahu. A jasno razvidno je, zakaj takšna inteligenca ne more biti transparentna ali popolnoma predvidljiva.[22] Pa vendar se, če ne zadostimo tema pogojema, znajdemo v situaciji inženirja, ki ga opisujeta avtorja: »No, ne vem, kako bo to letalo, ki sem ga izgradil, letelo varno – dejansko ne vem niti, kako bo sploh letelo, ali bo mahalo s krili, se napolnilo s helijem, ali pa na kak tretji način, ki si ga nisem niti predstavljal – a zagotavljam vam, da je zelo, zelo varno.«[23]

1.3 Sodobni izzivi

Zgoraj opisane dileme orišejo večplastnost etičnega izziva, ki ga predstavljata digitalizacija in razvoj umetne inteligence, a preden se lotimo odgovarjanja na zgoraj zastavljena vprašanja, si odgovorimo na najbolj pogosto javno vprašanje, povezano predvsem z robotiko in umetno inteligenco: »Če nas vse zamenjajo naprave, kaj bo potem z nami in delovnimi mesti?«

1.3.1 Avtomatizacija in delovna mesta

V laičnem odgovoru večinoma najdemo prepričanje, da avtomatizacija dela vodi v nezaposlenost, posledično pa v revščino, akumulacijo kapitala na vrhu socio-ekonomske prehranjevalne verige, in še večji prepad med bogatimi in revnimi. Acemoglu in Restrepo izpostavljata zmotno dihotomijo, v okviru katere obstajata samo dva možna odgovora na zgornje vprašanje; prvi, 'alarmistični' odgovor zaobjema omenjeno nezaposlenost in revščino, drugi pa zatrjuje, da avtomatizacija predstavlja odpiranje novih znanstvenih in tehničnih področij, s čemer prinaša tudi

[21] Ibidem.
[22] Ibidem.
[23] Ibidem, str. 5.

nova delovna mesta, in da razloga za skrb ni.[24] Acemoglu in Restrepo ugotavljata, da medtem, ko lahko način avtomatizacije, pri kateri je glavni cilj nadomeščanje človeškega dela, vodi v negativne posledice na področju trga dela in zaposljivosti, je v realnosti avtomatizacija počasnejši in nekoliko bolje reguliran proces, kot ga dojema večina javnosti. Skladno s tem lahko predpostavljamo, da bo izgubi delovnih mest v veliki večini primerov sledil proces nadomeščanja delovnih mest, ki odpira drugačna delovna mesta v podobni skupni količini. Te ugotovitve podpirata s podrobno matematično obravnavo in analizo preteklih trendov avtomatizacije in potrebe po človeškem delu. Problem, ki se utegne ohraniti, je ustrezno in pravočasno izobraževanje oziroma usposabljanje obstoječe delovne sile za nove potrebe na trgu dela.[25]

Akst vidi problem drugje. V svoji analizi preteklih posledic avtomatizacije v 20. stoletju ugotavlja, da je bil samo v ZDA upad števila redno zaposlenih moških med letoma 1960 in 2009 kar 18 odstotkov.[26] Zaključek njegovega članka se osredotoča na trditev, da je problem nezaposlenosti v luči avtomatizacije socio-političen, in ne tehnološki problem – namreč, da je tistim, ki jim avtomatizacija odvzame delovno mesto in (še) nimajo ustrezne izobrazbe ali znanj za prestop na drugo delovno mesto, ali pa drugih zaposlitvenih možnosti v njihovem sektorju enostavno ni, potrebno zagotoviti socialne storitve (zdravstveno zavarovanje, osnovni dohodek, ipd.) neodvisno od njihovega zaposlitvenega statusa.

Precej bolj optimistično stališče zagovarja Autor,[27] ki predvideva, da bodo mnoga delovna mesta ostala neavtomatizirana zaradi narave dela – npr. dela ki zahtevajo znanja iz mnogih področij na srednjem nivoju, in koordinacijo med temi znanj, ali pa dela, ki zahtevajo zmožnost odločanja in neposrednega dela z ljudmi. Podobno kot Acemoglu in Restrepo[28] tudi Autor izpostavja, da bodo potrebne izobraževalne reforme za zapolnjevanje novonastalih delovnih mest, a kot uspešen primer izpolnitve te zahteve iz preteklosti navaja reformo ZDA v začetku 20. stoletja, ko je ta država za potrebe novih delovnih mest kot prva na svetu uvedla univerzalno

[24] Acemoglu, D. in Restrepo, P.: Artificial Intelligence, Automation, and Work, v: Agarwal, A., Goldfarb, A. in Gans, J.: NBER Working Paper Series (24196), National Bureau of Economic Research, Cambridge 2018. (op. spletno dostopno brez oštevilčenih strani)
[25] Povzeto po: Acemoglu in Restrepo, Artificial Intelligence.
[26] Akst, D.: Automation Anxiety, v: Wilson Quarterly, 37 (2013) 3, str. 65-77.
[27] Autor, D. H.: Why Are There Still So Many Jobs? The History and Future of Workplace Automation, v: Journal of Economic Perspectives, 29 (2015) 3, str. 3-30.
[28] Glej: Acemoglu in Restrepo, Artificial Intelligence.

srednješolsko izobraževanje. Problem, ki po mnenju Autorja ostaja, če pride do masovne avtomatizacije je, kako razporediti kapital, ki bi ga prinesla višja produktivnost avtomatizirane industrije in storitev brez dodatnega stroška človeškega dela.

Tudi drugi avtorji na tem področju prihajajo do podrobnih zaključkov, pri čemer jih morda najbolje povzema ta zelo neposreden in iskren citat: »Razkrinkajmo torej lažno neizogibnost trenutne smeri kapitalističnega razvoja in si dalje predstavljajmo različne odnose med tehnologijami, zaposlitvijo in izobrazbo – in naj to storimo skupaj, v dialogu, v upanju za izgradnjo sveta, v katerem bi radi živeli v prihodnosti.«[29]

K učinkom ekonomskega interesa na različne varnostne mehanizme, povezane z digitalizacijo in umetno inteligenco, se bomo še vrnili. Sedaj, ko smo razčistili primarno eksistencialno krizo avtomatizacije dela, se lahko osredotočimo na druge težave, ki se pojavljajo z napredkom tehnologije. Preden preidemo na takšne ali drugačne avtonomne mehanizme in umetne inteligence, je smiselno pregledati tudi problematiko digitalne tehnologije, s katero človek neposredno upravlja, in ki je že par desetletij dostopna splošni javnosti.

1.3.2 Spletna anonimnost in zasebnost

S tem je seveda mišljen internet in vse njegove raznolike uporabe, z možnimi zlorabami, ki spadajo zraven. Otroci interneta smo bili vzgojeni ob svaritvah glede t.i. pojava »cyber-bullying«, spletnega nadlegovanja, sistemsko izkoriščanje razsežnosti interneta v kriminalne, politične in ostale namene, pa postaja jasno šele preteklo desetletje. Knjiga Bernarda E. Harcourta z naslovom *Exposed: Desire and Disobedience in the Digital Age* (2015) postreže s pompozno zvenečimi naslovi poglavij, ki izhajajo iz književnih del znanstvenofantastične distopije, od Velikega Brata do Panoptikona, pa od Mrka humanizma do Jeklene mreže.[30] Na prvi pogled se analogija med sodobno spletno družbo in zgornjimi koncepti zdi pretirana, a

[29] Peters, M. A., Means, A. J., in Jandrić, P.: Introduction: Technological Unemployment and the Future of Work, v: Peters, M. A., Means, A. J., in Jandrić, P.: Education and Technological Unemployment, Springer, Singapore 2019, str. 11.
[30] Harcourt, B. E.: Desire and Disobedience in the Digital Age, Harvard University Press, Cambridge 2015.

raziskave kažejo, da je najbolj priljubljena izmed vseh digitalnih tehnologij te črnoglede scenarije tesno približala resničnosti.[31]

Tehnologija, o kateri govorimo, so seveda socialna omrežja, ki so v preteklih 15 letih uspešno nadomestila tradicionalne medije, in premaknila javno sfero iz fizičnega v digitalni svet. Balkin socialna omrežja identificira kot eno izmed treh kategorij internetnih storitev – prva so osnovni sistemi (DNS, 'caching', itd.), druga pa plačilne storitve – ki ima svoj namen.[32] Slednji se ponovno deli na tri splošne namene, ki so olajšanje javnega sodelovanja, organizacija javnega dialoga in kuracija javnega mnenja. Kuracija javnega mnenja izhaja iz notranjih pravil in standardov posameznih socialnih omrežij, ki lahko omejijo vsebino objav po lastni izbiri. Relevanca te kuracije nadalje stremi iz prepričanja, da bi morale vlade vzdrževati omrežno nevtralnost (ang. 'Net Neutrality'), ker jim to preprečujejo ustave večine držav – z državno kuracijo interneta bi namreč prišlo do ekstenzivnih kršitev določenih temeljnih pravic, npr. svobode izražanja in svobodnega dostopa do informacij.[33]

Notranja politika socialnih omrežij mora biti torej tista, ki vzdržuje njihove družbene vloge, preprečuje širjenje lažnih podatkov, sovražnega govora, in podobnega. Balkin del rešitve vidi v pluralnosti institucij (tj. socialnih omrežij in medijev), ki stojijo med državo in posameznikom; te naj bi zagotavljale, da se uporabniki držijo standardov spoštljivega in korektnega dialoga v skladu z njihovimi standardi, ob predpogoju, da imajo tovrstne standarde zaradi njihovega 'dobrega imena' in socialnega statusa. Drug del rešitve se utegne nahajati v raznovrstnosti mnenj in vrednot, ki se pojavljajo v javni sferi.[34] Dodajmo, da je smiseln pogoj, da so ta mnenja in vrednote enakovredno dostopne in podprte z določenimi objektivnimi premisami, v nasprotnem primeru pa je odgovornost uporabnika, da jih zavrne oz. obravnava kritično.

Do zapletov pride, ko se pojavi zahteva po človeški moderaciji. Dosledna moderacija namreč zahteva povečano število moderatorjev, ki so bolje kvalificirani za objektivno kritično presojo vsebin. Ker je večina socialnih omrežij usmerjena proti

[31] Primer: Mozur, P. in Krolik, A.: A Surveillance Net Blankets China's Cities, Giving Police Vast Powers, v: The New York Times, 17. 12. 2019.
[32] Balkin, J. M.: How to Regulate (and Not Regulate) Social Media. Uvodni nagovor simpozija Association for Computing Machinery Symposium on Computer Science and Law, New York 2019.
[33] Povzeto po: Balkin, How to Regulate.
[34] Ibidem.

temu, da služijo z oglaševanjem čimvečji populaciji uporabnikov, je takšna poteza v nasprotju z njihovimi interesi. Je namreč že inherentno dražja, kot tudi lahko omejitve svobode izražanja pomenijo upad števila uporabnikov. Poleg tega socialna omrežja njihovi ustvarjalci pogosto dojemajo kot zabavo, profitno dejavnost, ali nekaj tretjega, prej kot pa jih dojemajo v okviru zgoraj opisanih družbenih namenov. Pravilna samorefleksija in sprememba oz. prilagoditev ekonomskega modela sta torej ključnega pomena za uspešno regulacijo socialnih omrežij.[35]

Tudi prvi dve internetni storitvi po Balkinu nista varni brez tveganj. Pri storitvah DNS in storitvah spletnega bančništva se identiteta uporabnika preverja z digitalnimi certifikati; mehanizmi, ki varujejo zaupne osebne podatke uporabnika. Uporabnik je v tem primeru t.i. 'digitalna avtoriteta' nad svojimi certifikati. Težava je, da je s človeškim uporabnikom moč manipulirati, da izda podatke certifikata, ali pa njegov certifikat poneveriti z vdorom v njegov (navadno manj zaščiten) osebni sistem. Pravno-politične rešitve so proti temu problemu praktično nemočne, saj je sledenje digitalnemu zločinu mnogo težje, kot razkritje fizičnega. Na srečo računalniška znanost v zadnjih letih predstavlja vse uspešnejše zaščitne protokole, npr. CT (transparentnost certifikata) in SK (neodvisni ključ).[36]

Vrnimo se k socialnim omrežjem in naslednji težavi, ki se pri njih ponavlja – anonimnosti. Različne spletne platforme imajo različne stopnje anonimnosti. Identiteta posameznika na globokem spletu je skoraj neizsledljiva, na nekaterih omrežjih je dobro zakrita, na nekaterih je kljub psevdonimu odkriti identiteto uporabnika povsem enostavno, če ta ni zelo previden pri ustvarjanju profila. Varnost na vseh izmed njih je stvar računalništva, in ne etike, zato se vanje ne bomo poglabljali.

Nekatera omrežja, kot je recimo Facebook, pa anonimnost kratkomalo prepovedujejo. Facebook ima strogo politiko resničnega imena, in moderatorji zaposleni pri tem omrežju aktivno iščejo profile, skrite za psevdonimom, in jih odstranjujejo z omrežja. Prednosti in slabosti te politike so stvar debate. Tisti ki jo zagovarjajo trdijo, da anonimnost vodi k širjenju lažnih informacij in sovražnega govora, ker brez identifikacije uporabnik nima strahu pred posledicami; po drugi

[35] Ibidem.
[36] Povzeto po: Laurie, B. in Doctorow, C.: Secure the Internet, v: Nature, 491 (2012), str. 325-326.

strani zagovorniki anonimnosti trdijo, da anonimnost vodi v večjo iskrenost, prav tako zaradi pomanjkanja strahu pred posledicami, ki pa vodi v večjo legitimnost podatkov.[37]

Resnica je, da obe trditvi držita, odvisni pa nista od anonimnosti same, ampak od namena, zaradi katerega se uporabnik odloči za anonimno delovanje. Bodle zadevo dojema kot stvar normativne etike.[38] Meni, da je stališče proti anonimnosti politično in ekonomsko utilitaristično, saj omogoča lažji nadzor nad prepričanji ne-anonimnih posameznikov; sam zagovarja nasprotno stališče, da je potrebno o anonimnosti odločati deontološko, se pravi z mislijo na pravice uporabnika, in ne samo na posledice njegove anonimnosti za tretje stranke.[39]

Tretji razvidni problem v sklopu socialnih omrežij in z njimi povezane digitalizacije pa je ta, ki to temo poveže z jedrom tega eseja – tj. z umetno inteligenco. Na socialnih omrežjih se namesto človeških moderatorjev vse pogosteje uporabljajo enostavne umetne inteligence oz. algoritmi, ki skrbijo za notranje in zunanje oglaševanje omrežja, nadzor nad objavami, komentarji, prijavami itd. Brkan izpostavlja štiri glavne težave, ki jih povzročajo ti inteligentni algoritmi – prva izmed teh je personalizacija, se pravi mehanizem oglaševanja oseb, izdelkov, političnih strank, in še česa, na podlagi zgodovine iskanja, objav, komentarjev in 'všečkov' na uporabnikovem profilu.[40] Ko govorimo o političnem in ideološkem oglaševanju, je personalizacija glavni promotor lažnih novic in pristranskosti, saj je ena od njenih največjih pomanjkljivosti učinek mehurčka (ang. Filter bubble effect). To pomeni, da se uporabnik nahaja med oglasi, ki so ustvarjeni specifično zanj, tako da potrjujejo njegova prepričanja in pred njim skrivajo alternativne možnosti. To je pogosto v volitvenih kampanijah, kjer lahko s pomočjo tega učinka volivca z oglasi polarizirajo (tj. neopredeljenega volivca usmerijo proti določeni odločitvi) in manipulirajo (tj. spremenijo njegovo politično usmerjenost na podlagi afirmacije drugih prepričanj).[41] Seveda obstajajo možne rešitve, najenostavnejša izmed katerih je odgovornost

[37] Povzeto po: Bodle, R.: The ethics of online anonymity or Zuckerberg vs. 'Moot', v: ACM SIGCAS Computers and Society, 43 (2013) 1, str. 22-35.
[38] Bodle, Zuckerberg vs. 'Moot', str. 23.
[39] Ibidem.
[40] Brkan, M.: Freedom of Expression and Artificial Intelligence: on personalisation, disinformation and (lack of) horizontal effect of the Charter, v: MCEL Working Paper Series, Maastricht 2019, str. 1-18.
[41] Povzeto po: Brkan, Freedom of Expresson; Burkell J. in Regan, P. M.: Voter preferences, voter manipulation, voter analytics: policy options for less surveillance and more auntonomy, v: Internet Policy Review, 8 (2019) 4, str. 1-24.

volivca, da razišče alternative, a je enostavno preveč naivna, da bi bila izvedljiva. V poštev prideta še odgovornost socialne platforme, v skladu s katero bi bila dolžna takšne oglase omejiti ali pa razkriti njihovega naročnika, ter sodna intervencija. Če pobrskamo nekaj strani nazaj po tem eseju, ugotovimo, da marsikomu to ni v pretiranem interesu.[42]

Težavi z avtomatizirano moderacijo socialnih omrežij sta tudi avtomatsko blokiranje/odstranjevanje nezakonitih vsebin in democija škodljivih (a legalnih) vsebin. Oba izmed teh principov sta sporna zaradi visoke zmotljivosti inteligentnega algoritma, ki bi naj o tem odločal, ter lahko kršita pravico svobode izražanja. Še ena težava se pojavi pri pravici do pozable (ang. Right to be forgotten),[43] v skladu s katero mora omrežje izbrisati vse podatke ne-anonimnega uporabnika na njegovo zahtevo. Ta lahko krši pravico do dostopa do informacij drugih uporabnikov.[44] Najbolj učinkovita rešitev opisanih težav bi bila bržkone poprej predlagana povečava števila in dvig zahtevanih kvalifikacij človeških uporabnikov, a ponovno, temu nasprotujejo ekonomski in politični interesi.

1.3.3 Algoritemska diskriminacija

Dodaten primer nevarnosti pomanjkanja internetne anonimnosti je prenos človeških predsodkov na spletne platforme. Sam izumitelj interneta, Sir Berners-Lee, je letos ponovno opozoril na diskriminacijo, ki se na spletu pojavlja proti ženskam, LGBT skupnosti, in še marsikomu, ter vključuje med drugim izsiljevanje, grožnje in spolno nadlegovanje.[45] A diskriminacija ne ostaja omejena na medčloveški odnos; človeški faktor v programiranju oglaševalnih in drugih algoritmov je diskriminacijo prenesel tudi v avtomatizirane procese, ki jih izvaja umetna inteligenca, učena iz človeških učnih podatkov.

Ti procesi lahko privedejo do različnih neželenih učinkov. Blažji so, da program za iskanje letalskih kart priporoči prej dražje karte glede na državo, katere državljan ga uporablja, ali pa da ob iskanju zdravstvenih nasvetov Google prej prikaže laične, pogosto nezanesljive spletne strani, kot pa strokovne. Med težjimi se dogaja, da

[42] Povzeto po: Balkin, How to Regulate; Burkell in Regan, Voter preferences.
[43] C-131/12, *Google Spain,* ECLI:EU:C:2014:317.
[44] Povzeto po: Brkan, Freedom of Expression.
[45] Povzeto po: Sample, I.: Internet 'is not working for women and girls', says Berners-Lee, v: The Guardian, 12. 3. 2020.

lahko program, ki izbira kandidate na razpisu za delo, na podlagi algoritemske diskriminacije nepošteno izloči pripadnike manjšin; program, ki dodeljuje kredite, lahko zavrne kredit neprimerljivo večjemu številu žensk kot moških. Naštejemo lahko na stotine identificiranih primerov tovrstne diskriminacije.[46][47]

Hacker trdi, da do algoritemske diskriminacije pride bodisi zaradi pristranskih učnih podatkov, ali pa zaradi neenakopravne temeljne resnice.[48] Pristranski učni podatki lahko nastanejo kot posledica nepravilnega ravnanja s podatki, ki izhaja iz pristranskosti oz. predsodkov tistega, ki te podatke pripravlja, ali pa zaradi napačne reprezentacije podatkov v smislu neenakomerno razporejenega vzorca, na katerem se umetna inteligenca uči. Neenakopravna temeljna resnica izhaja iz statistične diskriminacije, tj. iz splošnih podatkov, ki govorijo proti ali za določeno demografsko skupino.[49] Primer statistične diskriminacije je označitev afriških američanov kot bolj verjetnih, da storijo kriminalno dejanje. Čeprav je ta statistika do neke mere resnična, vseeno temelji na globlje zakoreninjeni človeški diskriminaciji, zaradi katere je bolj verjetno, da bo afriški američan obtožen, obsojen, obsojen na strožjo kazen, ali celo po krivem obsojen kot bel američan, in je zato, seveda, diskriminatorna.

Algoritemska diskriminacija lahko vodi tako v neposredno diskriminacijo, usmerjeno proti posamezniku na podlagi demografske skupine, ki ji pripada, ali pa v posredno diskriminacijo, pri kateri gre za diskriminacijo proti določeni demografski skupini, ki jo navadno povzroča navidez nevtralen algoritemski parameter. Identifikacija slednje je izjemno problematična, saj se pogosto niti ustvarjalci algoritma, niti njegovi uporabniki ne zavedajo tovrstnih težav.[50]

[46] Povzeto po: Hacker, P.: Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law, v: Common Market Law Review, 55 (2017), str. 1143-1186.
[47] Povzeto po: Sandvig, C. in ostali: Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, v: Data and Discrimination: Converting Critical Concerns into Productive Inquiry, 64th Annual Meeting of the International Communication Assiociation, 2014.
[48] Hacker, Teaching Fairness, str. 1147.
[49] Povzeto po: Hacker, Teaching Fairness.
[50] Ibidem.

Prvi korak do preprečitve algoritemske diskriminacije je odkritje diskriminacije, kjer se ta pojavlja, kar je mogoče z revizijami delovanja algoritmov.[51] Sandvig in ostali[52] trdijo, da so revizijske študije, ki preverjajo diskriminacijo tako, da preverjajo ljudi vpletene v postopek okoli algoritma, pogosto nesmiselne, če so subjekti teh študij obveščeni o njih, ali pa neetične, kadar niso, saj kršijo pravico subjektov do seznanjenosti o študiji, v kateri sodelujejo. Avtor zato predlaga, da se revidira algoritme same, kar se da neposredno - v ta namen predlaga 5 metod revizije. Prva izmed njih zahteva razkritje celotne kode algoritma v pregled, druga preučuje uporabnike oz. rezultate, ki so jih prejeli, ter preverja delovanje algoritma skozi rezultate glede na demografsko pripadnost uporabnikov, tretja pošilja umetno ustvarjene profile oseb algoritmu, in preučuje odgovore glede na parametre lažnih profilov, četrta uporabi resnične osebe z različnih ozadij, ki sodelujejo z revizorji in z njimi delijo svoje rezultate, peta pa uporabi rezultate čimvečjega števila pripadnikov čimbolj raznolike populacije, ki je že pred obveščenostjo o raziskavi prejela rezultate preučevanega algoritma.[53]

Naslednja stopnja po odkritju diskriminacije je odgovorno oz. pošteno pridobivanje podatkov, najpomembneje učnih podatkov za algoritme, ki jih lahko pridobivamo s pred-, med-, ali post-procesiranjem. Pred-procesiranje zahteva natančen pregled in po potrebi prilagoditev izhodiščnih podatkov, v kolikor so ti oporečni; med-procesiranje predvideva anti diskriminacijske elemente v samem algoritmu, ki pridobiva podatke; post-procesiranje pa modifikacijo pridobljenih podatkov.[54] V skladu z zakonodajo proti diskriminaciji se lahko tovrstne tehnične rešitve pri algoritemski obravnavi ljudi tudi zakonsko predpišejo.

1.4 Umetna inteligenca danes

1.4.1 Zanesljivost umetne inteligence

Pri obravnavi takšnih primerov in z mislijo na nadaljnji razvoj umetne inteligence, se človek vpraša, do kakšne mere lahko zaupamo tehnologiji, ki temelji na nas, vključno z vsemi našimi napakami, pa vendar ohranja določeno mero avtonomnosti,

[51] Hajian, S. in ostali: Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining, v: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 22 (2016), str. 2125-2126.
[52] Sandvig i.o., Auditing Algorithms.
[53] Ibidem.
[54] Povzeto po: Hajian i.o., Algorithmic Bias.

zaradi katere nam je vpogled v njeno podrobno delovanje zelo otežen, če ne že kar onemogočen.

Pogosto ima zanesljivost delovanja umetne inteligence dve plati: raziskava potencialne uporabe UI v človeških misijah na Luno ali celo Mars je pokazala, da medtem, ko so ti sistemi zanesljivi pri nadomeščanju človeka v namene prognostike ostalih sistemov (predvidevanje okvar, možnih ukrepov, itd.), kompleksnost same UI onemogoča človeškim delavcem, da bi pred uporabo ustrezno verificirali njeno delovanje.[55] To je ilustracija omenjene dvostranskosti. Večinoma ima umetna inteligenca pred človekom prednost pri samih nalogah, ki jih izvaja: UI se je izkazala pri medicinski diagnostiki, saj ima možnost hitrega procesiranja velike količine podatkov in iskanju ustrezne diagnoze na podlagi preteklih primerov; sonde, roverji in podobne naprave, opremljene z UI so se zelo dobro obnesle v vesolju in globinah oceanov zaradi svoje neobčutljivosti na zunanje pogoje; v drugih primerih je UI prekosila človeka zaradi svojega ne-čustvenega, popolnoma logičnega in objektivnega odločanja. Istočasno UI predstavlja težavo zaradi netransparentnosti in nečloveškosti: kompleksen sistem UI, ki se lahko hitro uči in dela popravke, ni transparenten za človeški pregled v njegovo delovanje, in lahko zelo hitro uide izpod nadzora; UI, ki takšnega učenja ni zmožnja, ne more predvideti ali se pravilno odzvati na nepredvidljive okoliščine, prav tako pa se ne more naučiti kompleksnih pogojnih sistemov, npr. etike.[56]

1.4.2 Samovozeča vozila

Na nobenem sodobnem področju umetne inteligence te pomanjkljivosti ne pridejo tako dobro do izraza kot pri samovozečih vozilih, ki se v zadnjem desetletju vse hitreje premikajo iz sfere teoretične ali eksperimentalne tehnologije v naša vsakdanja življenja. Večina sveta rabo avtomatiziranih vozil še zmeraj strogo omejuje, čeprav obstajajo izjeme, npr. ameriška zvezna država Kalifornija, ki dopušča rabo samovozečih naprav brez potnikov pod določeno maso in najvišjo hitrostjo.

[55] Povzeto po: Schwabacher, M. in Goebel, K.: A survey of artificial intelligence for prognostics, v: AAAI Fall Symposium – Technical Report, 2007, str. 107-114.
[56] Povzeto po: Dasoriya, R. in ostali: The Uncertain Future of Artificial Intelligence, v: 8th International Conference on Cloud Computing, Data Science & Engineering, 2018, str. 458-461.

Medtem, ko je osnovna tehnologija za samovozečimi vozili, ki zaobjema sledenje cestno-prometnim predpisom, navigiranje med ostalimi vozili v normalnih okoliščinah, ter podobne aspekte njihove uporabe že praktično dovršena, pa je obnašanje samovozečih vozil v abnormalnih okoliščinah, npr. v primeru ekstremnega vremena, neizogibne nesreče, ali celo neizogibne smrti udeleženca ali udeležencev v prometu, še zmeraj neodgovorjeno in domala pereče vprašanje.

Pogost laičen odgovor na to, kako naj umetna inteligenca, ki upravlja vozilo, ravna v primeru neizogibne nesreče, je popolnoma utilitaristično: ravna naj tako, da se čimbolj zmanjša število žrtev oz. negativnih posledic na splošno. Čeprav to sprva zveni smiselno, gre v bistvu za globlji etični problem, ki se nahaja v neposredni analogiji s klasičnim filozofskim problemom vagona (ang. Trolley problem), kjer mora akter sprejeti odločitev, ali bo vagon, ki drvi proti petim ljudem, privezanim na tirih, preusmeril na drugi tir, na katerega je privezana samo ena oseba. Če ga ne preusmeri, je pasivno sodeloval pri smrti petih ljudi; če ga, je aktivno kriv smrti ene osebe. Klasična oblika problema ima še mnoge dodatne, kompleksnejše permutacije – kot jih ima tudi problem odločanja pri samovozečih vozilih.

De Sio[57] navaja, da je utilitaristično odgovor na dilemo ravnanja v neizogibnih nesrečah pri samovozečih vozilih nezadosten, saj je v nasprotju z bolj absolutnimi etičnimi normativami, najpomembneje z absolutno deontološko prepovedjo odvzema življenja nedolžne osebe, ki je zakoreninjena tudi v zakonodaji celotnega civiliziranega sveta. Vendar pa obstajajo izjeme, ki niso vedno popolnoma jasne. Omenjeni avtor ilustrira variabilnost principa z dvema razvpitima primeroma. Prvi je primer Dudleya in Stephensa, ki sta po večtedenskem stradanju po brodolomu ubila in pojedla tretjo osebo, ki se je z njima nahajala na območju brodoloma – po vrnitvi v ZDA sta bila obsojena umora.[58] Drugi primer sta Gracie in Rosie, siamski dvojčici, ki ju je bilo potrebno razdružiti:[59] Gracie je držala Rosie pri življenju, a so ocenili, da bosta skupaj zdržali le približno 6 mesecev. Po drugi strani bi kirurška ločitev Rosie brez dvoma ubila, medtem ko so ocenili, da ima Gracie 94% možnost preživetja. Sodišče je v tem primeru odločilo, da se ločitev izvede in da je smrt Rosie v tem primeru nujna za preživetje Gracie. Lahko bi trdili, da je bila v prvem primeru

[57] de Sio, F. S.: Killing by Autonomous Vehicles and the Legal Doctrine of Necessity, v: Ethic Theory and Moral Practice, 20 (2017), str. 411-429.
[58] R v Dudley and Stephens (1884) 14 QBD 273 DC.
[59] Re A (conjoined twins) [2001] 2 WLR 480.

smrt tretje osebe nujna za preživetje Dudleya in Stephensa, a sta zločina neposrednega umora in kanibalizma pripeljala do njune obsodbe. Nasprotno, namen operacije ni bil ubiti Rosie, da bi ohranili Gracie pri življenju, ampak formalno zgolj ločiti dvojčici. Sodstvo se je v tem primeru sklicevalo tudi na strokovno avtoriteto kirurgov, ki so o operaciji odločali.[60]

Drugi problem utilitaristične obravnave samovozečih vozil je inkomenzurabilnost človeškega življenja. Poenostavljeno, princip inkomenzurabilnosti predpostavlja, da ni takšnih parametrov, s pomočjo katerih bi bilo moč določiti katero življenje je vrednejše od drugega, pa naj poskušamo na podlagi starosti, spola, etnične pripadnosti, poklica, izobrazbe, ali katerekoli kombinacije teh ali drugih karakteristik oseb, udeleženih na primeru.[61] Vsaka tovrstna obravnava je inherentno diskriminatorna, poleg tega pa ne mora uskladiti razlik vrednotenja človeškega življenja, ki bi utegnile izhajati iz kulturnih razlik, npr. tradicionalne zahodnjaške cenitve mladega življenja nad življenjem starostnika v kontrastu s tradicionalno vzhodnjaško centivijo življenja starešine nad življenjem otroka.

Tretji problem je težavnost predvidevanja dolgoročnih posledic odločitve v posameznih primerih, kot tudi dolgoročnih posledic vzpostavitve enotne normativne etike za vsa samovozeča vozila.[62]

Thornton in ostali[63] predlagajo rešitev, ki vzpostavi kompatibilnost med normativnimi etikami. Njihov predlog izhaja iz strinjanja, da ena normativna etika ni zadostna za reševanje katerekoli posamezne situacije, v kateri se lahko znajde avtonomno vozilo. Po njihovem prepričanju mora vsako avtonomno vozilo zagotoviti trojici kriterijev, tj. mobilnosti, legalnosti in varnosti, pri čemer pogosto ni mogoče zadostiti vsem trem popolnoma. Primer, ki ga navajajo, je vozilo, ki je obtičalo za oviro na cesti. Na desni ima komaj kaj prostora, da se izogne oviri, kar bi ogrozilo varnost potnikov, na desni pa ga ovirata dve polni črti, ki mu preprečujeta prečkanje zaradi legalne zahteve, ki jo predstavljata. Če ne izbere nobene opcije, lahko obtiči za oviro za nedoločen čas, kar krši pravico do mobilnosti potnikov. V tovrstnih okoliščinah bi bilo, ob primerni preučitvi dodatnih varnostnih okoliščin,

[60] Povzeto po: de Sio, Doctrine of Necessity.
[61] Ibidem.
[62] Povzeto po: de Sio, Doctrine of Necessity.
[63] Thornton, S. in ostali: Incorporating Ethical Considerations Into Automated Vehicle Control, v: IEEE Transactions on Intelligent Transportation Systems, 48 (2017) 6, str. 1429-1439.

upravičeno prekršiti cestno-prometni predpis, ki prepoveduje prečkanje dvojne polne črte.[64]

Predlog, ki sledi, je konsekvencialistična etika, omejena z deontološkimi principi, ki izhajajo iz Asimovih treh načel robotike, ki med drugim zaobjemajo absolutno prepoved škodovanja človeškemu življenju, itd.[65] Ko se umetna inteligenca 'prepriča', da nobeno izmed deontoloških načel ne bo kršeno, lahko nadaljuje z ravnanjem v skladu s konsekvencialistično etiko, ki predvideva najmanj škodljive posledice dejanja.[66]

Vseeno pa tudi tovrstna etična shema ne predvideva resnično anomalnih situacij, kjer npr. ni mogoče ne-prekršiti deontološkega načela. Wagner in Koopman[67] trdita, da to izhaja iz razlik v človeškem ravnanju in ravnanju umetne inteligence, ki upravlja vozilo. Vozila namreč ravnajo v skladu s principom induktivne inference, torej lahko v skladu z opazovanji izvajajo izjemno natančne operacije, dokler te ustrezajo ustaljenemu vzorcu. Kadar mu ne, odločitev postane nemogoča. Kot rešitev predlagata metodo umetnega učenja, ki temelji na falsifikacionistični teoriji, v skladu s katero se lahko umetna inteligenca uči iz neuspešnih poskusov reševanja problemov, najbolje v simuliranih anomalnih okoliščinah.[68]

Četudi se sčasoma najde unificiran etični standard, ki bo samovozeča vozila naredil enako ali bolj zanesljiva kot človeške voznike, tudi v anomalnih pogojih, pa bo najbrž še vedno obstajalo tranzicijsko časovno območje med človeško vožnjo in popolno avtomatizacijo, kjer bo umetna inteligenca upravljala samo del procesa (avtopilot, avtomatsko zaviranje v sili, itd.). Kako pri različnih stopnjah avtomatizacije pripišemo krivdo za nezgodo, oz. primere nesreč zakonsko obravnavamo?

Anderson predvideva, da v popolnoma avtomatiziranih vozilih krivda voznika ne bo več veljaven faktor, ampak bo krivda avtomatsko na strani proizvajalca vozila, pri čemer lahko gre za napake v proizvodnji ali napake v načrtovanju.[69] Kazenska in

[64] Povzeto po: ibidem.
[65] Ibidem.
[66] Ibidem.
[67] Wagner, M. in Koopman, P.: A Philosophy for Developing Trust in Self-Driving Cars, v: Meyer, G. in Beiker, S.: Road Vehicle Automation, Springer, New York 2015, str. 163-171.
[68] Povzeto po: Ibidem.
[69] Anderson, J. M. in ostali: Liability Implications of Autonomous Vehicle Technology, v: Anderson, J.M. in ostali: Autonomous Vehicle Technology, RAND Corporation, Santa Monica 2014, str. 111-134.

odškodninska odgovornost proizvajalca sta pri tem odvisni od različnih načinov zakonske obravnave, ki jih bomo v namene tega prispevka izpustili. Avtomatizirana vožnja bo neizogibno privedla tudi do nove vrste nesreč, kot so nesreče zaradi napak v programiranju, ki bi ga lahko vsaj delno izvajal tudi uporabnik, ne le proizvajalec avtomobila, ali nesreče delno- ali ne-avtomatiziranih vozil in pešcev, navajenih na avtomatizirana vozila, na katera ne rabijo biti pozorni. V prvem primeru je lahko krivda na strani uporabnika, če ta ni upošteval navodil programiranja, ali pa proizvajalca, če navodila niso bila ustrezna; v drugem primeru je lahko kriv voznik, ki ni videl pešca, ali pa pešec, ki je ravnal neodgovorno ob pričakovanju avtomatiziranega vozila.[70]

Pri delno-avtomatiziranih vozilih ni vprašanje odgovornosti nič manj kompleksno, pri vozilu z avtocestnim avtopilotom je lahko kriv voznik, ki je med avtomatsko vožno zadremal proti opozorilom proizvajalca, ali pa proizvajalec, ki ni ustrezno opozoril na to, da umetna inteligenca v vozilu še zmeraj zahteva človeški nadzor.[71] Bellet in ostali[72] predstavljajo kompleksno shemo devetih faz HMT (Human-Machine Transition; Prehod med človekom in napravo), ki zaobjema pravne podrobnosti, v katere se ne bomo podrobneje spuščali. Mnogo enostavneje, Bryson in Winfield[73] ponovno poudarjata zahtevo po transparentnosti programske opreme, ki vodi umetno inteligenco v vozilu, tako za uporabnika, v smislu vpogleda v to, kako bo njegovo samovozeče vozilo ravnalo in zakaj, kot za pristojne službe, ki bi utegnile preizkovati morebitno nezgodo ali podobne okoliščine, v smislu črne skrinjice na letalih.

1.4.3 Avtomatizirano orožje

Še večjo etično dilemo predstavljajo drugi avtomatizirani sistemi, še najbolj izmed njih avtomatizirano orožje. V vojaških konfliktih vsebolj narašča uporaba brezpilotnih zrakoplovov oz. dronov, ki se jih vodi ali pa nadzoruje na daljavo. Že pri samih dronih imajo mnogi pomisleke glede moralnih aspektov njihove rabe proti

[70] Povzeto po: Ibidem.
[71] Ibidem.
[72] Bellet, T. in ostali: From semi to fully autonomous vehicles: New emerging risks and ethico-legal challenges for human-machine interactions, v: Transportation Research, 63 (2019) F, str. 153-164.
[73] Bryson, J. in Winfield, A.: Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems, v: Computer, 50 (2017) 5, str. 116-119.

ljudem; mnogi jim očitajo neosebnost, kršenje dostojanstva ubitega sovražnika, in pripisujejo večjo kolateralno škodo kot tradicionalnim orožjem.

Statman trdi, da na primeru dronov to ne drži. [74] Glavna uporaba dronov je namreč v skladu s principom nepotrebnega tveganja, saj lahko en sam dron nadomesti zajetno število živih vojakov, ki jim zaradi uporabe drona ni potrebno tvegati življenja na bojnem polju. Nadalje Statman nasprotuje tudi tistim, ki trdijo, da droni povzročajo večjo kolateralno škodo, pri čemer se opira na njihovo natančnost v primerjavi z drugimi modernimi načini vojskovanja, npr. bombnimi letali, in tistimi, ki trdijo, da gre za kršitev dostojanstvene smrti in razčlovečenje, rekoč, da dron ni nič manj neoseben kot bojna konica, izstreljena z vojaškega letala, ali pa ostrostrelec, katerega prisotnosti se tarča sploh ne zaveda. Potrebno je poudariti, da ti argumenti veljajo samo, kadar je konflikt v vsakem primeru neizogiben, in bi se v primeru prepovedi/nezmožnosti uporabe dronov uporabila druga orožja.[75]

Vseeno pa njegov zagovor vodenih, delno-avtomatiziranih sistemov, ne nudi zadostnega odgovora na vprašanje polno avtomatiziranih sistemov, ki jih vodi umetna inteligenca. Swoboda[76] preučuje problematiko odgovornosti pri polno avtomatiziranih orožjih; tovrstna orožja naj bi se učila iz lastnih izkušenj, na podlagi vnaprej programiranih parametrov za učenje. Pri tem je težavno, da osnovni program takšnega 'vojnega robota' ne more nikdar vsebovati vseh možnih okoliščin, zaradi česar bo njegova reakcija v novih okoliščinah nepredvidljiva. Programer za to ne more bit odgovoren, saj ni mogel predvideti izredne okoliščine, robot pa prav tako ne, vsaj zaenkrat še ne, ker ni oseba, ki bi lahko bila nosilka pravic in dolžnosti. Večino izrednih scenarijev je mogoče rešiti s programiranjem principa pozitivne diskriminacije, na podlagi česar bi lahko robot razločeval med sovražnikom, nevtralnim subjektom in prijateljem ne glede na druge okoliščine, a tudi tukaj obstajajo izjeme: sovražniki bi se lahko zakrinkali kot prijatelji, robot bi lahko prejel sporne direktive nadrejenih ljudi, nenazadnje tudi ni odporen proti programskim vdorom.[77]

---

[74] Statman, D.: Drones and Robots: On the Changing Practice of Warfare, v: Lazar, S. in Frowe, H.: The Oxford Handbook of Ethics and War, Oxford University Press, Oxford 2015. (op. spletna predizdaja poglavja brez oštevilčenih strani)
[75] Povzeto po: Ibidem.
[76] Swoboda, T.: Autonomous Weapon Systems – An Alleged Responsibility Gap, v: Müller, V. C..: Philosoph and Theory of Artificial Intelligence, Springer, Leeds 2017, str. 302-314.
[77] Povzeto po: Ibidem.

Ena izmed možnih rešitev (morebiti, po mnenju nekaterih tudi edina) za moralne dileme pri vseh avtonomnih sistemih bi bil razvoj človeku podobne umetne inteligence, ki bi etične dileme dejansko razumela, in se z njimi soočala kot človek, s čemer bi lahko postala tudi zakonsko in kazensko odgovorna za svoja dejanja. A takšna umetna inteligenca brez dvoma predstavlja še več novih etičnih in ontoloških vprašanj.

1.4.4 Razvoj človeku podobne umetne inteligence

Osnovna ontološka dilema je, kako bi sploh ustvarili umetno inteligenco, ki je podobna človeku – bržkone bi morala biti tudi fizično podobna človeku. Sodobna znanost si predstavlja nevrone v možganih kot omrežje logičnih vrat, kar jim omogoča izgradnjo t.i. nevronskih omrežij, na katerih temelji večina trenutne umetne inteligence. Izkazalo pa se je, da so umetna nevronska omrežja v mnogih karakteristikah popolnoma drugačna od možganov: imajo mnogo večje hitrosti (širjenja signalov), a je njihova zmožnost procesiranja smešno majhna; odlična so pri logičnih, racionalnih operacijah, kot so šah in matematika, a skoraj neuporabna v umetnosti, etiki, itd. Marsikdo bi to razhajanje pripisal razlikam v substratu, tj. v mehanski (baker-silikon) in biološki (ogljikovodiki) podlagi, morda bi morala imeti človeku podobna umetna inteligenca umetne možgane, zgrajene iz bioloških celic.[78] Seveda vse to pride v poštev šele, ko se podpišemo pod teorijo fizikalizma, ki zagovarja idejo, da je popolnoma vse zvedljivo na fizikalne pojave, vključno z zavestjo. Medtem ko je teorija v naravoslovnih znanostih široko sprejeta kot edina znanstveno korektna razlaga, filozofija opaža znatno razlagalno vrzel med možgani in zavestjo, nobena znanost namreč (še) ne zna razložiti vzročnosti med biološkim substratom in psihološkim fenomenom zavesti, ali pojasniti specifične 'takšnosti' izkustev, čustev, in podobnih subjektivnih – ter izjemno izrazito človeških – pojavov.[79]

Četudi se utegne fizikalizem z veliko verjetnostjo izkazati za veljavnega, in metoda za vzpostavitev človeku podobne umetne inteligence razvita, še zmeraj obstajajo etične dileme pri razvoju tovrstne inteligence. Medtem, ko so nekateri zadovoljni s

[78] Povzeto po: Brooks, R. in ostali: Is the Brain a Good Model for Artificial Intelligence, v: Nature, 482 (2012), str. 462-463.
[79] Povzeto po: Tye, M.: Ten Problems of Consciousness: A Representational Theory of a Phenomenal Mind. MIT Press, Cambridge 1995.

strogimi nadzornimi pogoji pri razvoju te tehnologije, zahtevajoč med drugim oziranje na kontekstualno delovanje umetnih 'oseb', spoštovanje inteligence kot domene človeškega, ter previdno obravnavanje vsake faze v nastajanju nove tehnologije,[80] pa drugi trdijo, da obstajajo razlogi, zakaj človeku podobne umetne inteligence sploh ne bi smeli razvijati. Poglejmo si na primer sledeč argument: v kolikor je umetna inteligenca zmožna replicirati vsa človeška čustva in izkustva, je zmožna potemtakem tudi trpeti, saj je trpljenje človeško izkustvo. Trpljenje je seveda inherentno slabo in ga je potrebno, vedno kadar je to mogoče, preprečiti. Človeško trpljenje že obstaja, torej ga ni mogoče preprečiti, tudi v prihodnje pa ga ni nujno mogoče preprečiti, zaradi česar moramo preprečiti samo posamezne primere človeškega trpljenja, ne pa tega kot celote; v primeru človeku podobne umetne inteligence, ki še ne obstaja, in bo zagotovo trpela kot posledica svoje človeškosti, je njeno trpljenje mogoče in, v skladu s teorijo antinatalizma, nujno potrebno preprečiti.[81]

Antinatalistična prepričanja zaenkrat stojijo na precej trhlih argumentacijskih temeljih, tako da se raje osredotočimo na scenarij previdnega, odgovornega razvoja umetne inteligence. Kot rečeno bo umetna inteligenca podobna človeku takrat, ko doseže zmožnosti zaznavanja, čutenja, in intencionalnosti. Kljub temu pa Kane[82] trdi, da so že precej nižje umetne inteligence lahko smatrane kot osebe. Njegov argument temelji na Heideggrovi shemi bivanja v svetu, kjer lahko bitnosti v svetu samo bivajo, lahko bivajo kot orodja, lahko pa bivajo-v-svetu (nem. Dasein), kar pomeni, da lahko spoznavajo svet in druge bitnosti v njem, ter z njimi vzpostavijo eksistencialne odnose. Kane smatra, da so mnogi obstoječi, učeči-se algoritmi, kot je Facebook for Politics ali DeepMind, zaradi svoje zmožnosti spoznavanja sveta, že Algoritemske umetne osebe (ALAP).[83]

[80] Povzeto po: Boddington, P.: Towards a Code of Ethics in Artificial Intelligence, v: Delphi, 2 (2019) 2, str. 105-106.
[81] Povzeto po: Beckers, S.: AAAI: An Argument Against Artificial Intelligence, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 235-247.
[82] Kane, T. B.: A Framework for Exploring Intelligent Artificial Personhood, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 255-258.
[83] Povzeto po: Kane, A Framework.

1.5 Umetna inteligenca v prihodnosti

1.5.1 Umetna superinteligenca

Ultimat razvoja umetne inteligence je po interpretaciji mnogih superinteligenca oz. singularna UI. Osnovna superinteligenca se splošno definira kot umetna inteligenca, ki je v vseh nalogah, ki jih zmore opravljati človek, vsaj za nek nezanemarljiv odstotek boljša od človeka. DeepBlue, šahovska UI, tako ni superinteligenca, ker je od človeka boljša samo v eni specifični nalogi. Singularna UI je teoretična superinteligenca, ki ima na razpolago kapacitete na redu kvantnega računalnika, in lahko z neprimerljivo višjo hitrostjo in učinkovitostjo opravlja človeške naloge. Takšne inteligence so posebej problematične, saj se lahko pojavijo nenadno, celo pomotoma zaradi pomanjkanja razumevanja kakšnega inovativnega novega nevronskega omrežja, imajo možnost izjemno hitrega učenja, replikacije, in opravljanja postopkov in operacij, ki bi lahko bile krepko izven dometa človeškega razumevanja, predvsem zaradi razlik v človeškem načinu razmišljanja in psihi, ter načinu razmišljanja in 'psihi' umetne inteligence. Superinteligence bi lahko bile nevarne zaradi potencialnega izkoriščanja – predstavljajmo si umetno superinteligenco, ki služi samo nekaj najbogatejšim in najvplivnejšim ljudem na svetu, bodisi same po sebi.[84]

Kako bi lahko bile nevarne same po sebi najbolje razloži miselni eksperiment, poimenovan Rokov Bazilisk, ki ga je leta 2010 na spletnem forumu Less Wrong objavil anonimni uporabnik Roko. Gre za reinterpretacijo Yudkowskyjevega koncepta CEV (Coherent Extrapolated Volition), ki predvideva superinteligenco, ki zastopa človeške interese v smislu izpeljave konvergentnih, množičnih interesov človeštva. Gre za t.i. 'prijazno umetno inteligenco', a eksperiment demonstrira, kako bi lahko šla zadeva narobe; UI z vprogramiranimi principi CEV še zmeraj ne razmišlja na enak način ko človek, in nima določenih človeških atributov, kot sta intuitivna etika in empatija. Interes, ki ga zastopa je zmanjšanje eksistencialne grožnje človeštvu, torej efektivno znižanje človeškega trpljenja na vse načine. Z zapleteno argumentacijsko shemo Roko svojo konceptualizacijo privede do možnega scenarija, da UI poskuša svoj cilj doseči s kaznovanjem tistih, ki so se cilja zavedali, a k njemu

[84] Povzeto po: Bostrom, N.: Ethical Issues in Advanced Artificial Intelligence, v: Schneider, S.: Science Fiction and Philosoph: From Time Travel to Superintelligence, Blackwell Publishing, Chichester 2009, str. 374-382.

niso prispevali, in nagrajevanju tistih, ki so vede prispevali. To ji uspe preko simulacije zavesti 'kaznjencev', a ker je zmožna popolne simulacije, je vsaka izmed simuliranih zavesti *de facto* človek, zaradi česar UI v iskanju svojega (dobronamernega) cilja (pomotoma) znatno zviša splošno raven trpljenja ljudi.[85]

Prinzing[86] predlaga alternativo modelu CEV in drugim dotedanjim modelom, ki temelji na učenju koncepta ljubezni umetnim inteligencam. Pri tem definira ljubezen kot odnost do neke osebe, pri katerem akter ravna v skladu z interesi tiste osebe neodvisno od vseh svojih interesov. UI bi tako v zgodnjih fazah učenja (preden doseže nivo superinteligence ali celo singularnosti) priučili takšno delovanje, a ne do posamezne osebe, temveč do celotnega človeštva. V primeru navzkrižja interesov med ljudmi ali frakcijami ljudi, bi morala takšna UI, zaradi enake ljubezni do vseh ljudi, zavzeti egalitarno stališče in se vzdržati sodelovanja v konfliktu, dokler ga ljudje ne rešijo sami brez njenega vmešavanja.[87] Potencialno bi lahko tak model preprečil tudi najbolj črnoglede scenarije kot je Rokov Bazilisk, a o tem ne moremo biti prepričani pred dejansko implementacijo kateregakoli etičnega modela.

Teoretizacije o superinteligenci in singularnosti morda zvenijo izjemno futuristične, če ne že kar nemogoče, a po anketiranem mnenju vodilnih znanstvenikov na področju UI, jih petdeset odstotkov meni, da bo UI dosegla človeku podobne lastnosti oz. človeški nivo delovanja do štiridesetih let tega stoletja, še nadaljnjih štirideset odstotkov pa, da bo ta nivo dosežen do leta 2075. Še več, kar peteminsedemdeset odstotkov jih je mnenja, da bo UI dosegla nivo superinteligence v roku največ tridesetih let od dosega človeškega nivoja delovanja. Na kratko: okoli 75% vodilnih raziskovalcev UI meni, da bo umetna superinteligenca postala resničnost do konca tega 21. stoletja. Čeprav jih večina trdi, da bo razvoj takšne UI za človeštvo v splošnem nekaj dobrega, jih 31% meni, da bo superinteligenca za človeštvo slaba oz. celo katastrofalna.[88]

[85] Povzeto po: <basilisk.neocities.org> (29. 5. 2020)
[86] Prinzing, M.: Friendly Superintelligent AI: All You Need Is Love, v: Müller, V. C.: Philosoph and Theory of Artificial Intelligence, Springer, Leeds 2017, str. 288-301.
[87] Povzeto po: Prinzing, Friendly Superintelligent AI.
[88] Povzeto po: Müller, V. C. in Bostrom, N.: Future Progress in Artificial Intelligence, a Survey of Expert Opinion, v: Müller, V. C.: Fundamental Issues of Artificial Intelligence, Springer, Berlin 2016, str. 553-571.

Umetna inteligenca na višjem nivoju bo še poglobila težave, s katerimi smo se soočali in se še soočamo v času avtomatizacije in digitalizacije; kakšna bo človeška funkcija v dobi, ko bo umetna inteligenca lahko delovala na ali nad človeškim nivojem zmožnosti, in to na vseh možnih področjih vključno z vzdrževanjem umetnih inteligenc samih? Za rešitev nastalega scenarija in preprečevanje distopičnega elitizma bo potreben resen nadzor nad postopkom razvoja UI, ter z njo povezanih etičnih standardov in zakonodaje, kot tudi obsežne etično-tehnične rešitve ki bodo preprečevale dominanco UI nad ljudmi in podobne zlorabe moči.[89]

1.5.2 Vprašanje umetnih oseb

Nazadnje se pri človeku-podobni UI pojavi še eno pereče vprašanje: kakšen bo status potencialno čuteče in empatične človeku podobne umetne osebe?

Mishra[90] identificira štiri različne kategorije, po katerih je lahko neko bitje kandidat za moralni status, ob prepoziciji da ima to bitje interes in je lahko v primeru kršitve tega interesa oškodovano. Prva kategorija je SCC (Sophisticated Cognitive Capacity – Prefinjena kognitivna zmožnost), druga kandidat za SCC (potencialno SCC v razvoju oz. ni določljivo, da ima SCC), tretja je posebni odnos (z bitjem z SCC, npr. domače živali), in RCC (Rudimentary Cognitive Capacities – Osnovne kognitivne zmožnosti). Mishra aplicira te kategorije na moralni status digitalnih (simuliranih) agentov, tukaj pa ga bomo uporabili za moralni status umetno inteligentnih oseb. Skoraj vsaka nekoliko razvita UI ustreza vsaj kategoriji RCC in ima osnovni moralni status, ki človeku preprečuje določene kršitve; človeku podobne umetne inteligence in superinteligence pa bi v celoti ustrezale kategoriji SCC.[91] V tem primeru bi jih bilo nemoralno zasužnjiti, torej jim mora biti dopuščena določena mera svobodne volje, kot tudi svoboda gibanja (moralo jim bo biti dodeljeno neke vrste telo). V primeru kršitev zakona bi jim morali soditi kot človeškim osebam, ter kazni ustrezno prilagoditi (izklop zavestne UI bi bil ekvivalenten usmrtitvi oz. umoru).

[89] Povzeto po: Wang, W. in Siau, K.: Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work, and Future of Humanity: A Review and Research Agenda, v: Journal of Database Management, 30 (2019) 1, str. 61-79.
[90] Mishra, A.: Moral Status of Digital Agents: Acting Under Uncertainty, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence, Springer, Leeds 2017, str. 273-287.
[91] Povzeto po: Ibidem.

Da se lahko sploh pogovarjamo o sodelovanju UI v družbi, pa morajo imeti dostop do podatkov, iz katerih se lahko učijo o tem, kaj je v človeški družbi sprejemljivo in kaj ne – potrebujejo torej dostop do institucionalnih dejstev. Preprosto učenje z imerzijo v človeško družbo se je že izkazalo za neefektivno na primeru Twitter robotke Tay, ki je v nekaj urah eksperimenta začela izkazovati rasistične, antisemitistične, in druge ksenofobne tendence.[92] Višje razvita umetna inteligenca bo morda imela zmožnost kritičnega mišljenja, s pomočjo katere bo takšne vplive lahko filtrirala, a tudi to bo moralo biti podprto s formalnimi institucionalnimi dejstvi človeške družbe. Prav tako bi bili smiselni mehanizmi, ki bi človeku preprečili indoktrinacijo učečih se umetnih inteligenc, podobno kot obstajajo ukrepi, ki pedagogom preprečujejo indoktrinacijo mladih učencev.[93]

1.6 Zaključek

Tekom tega stoletja se bo človeštvo soočilo z vse hitreje razvijajočo se umetno inteligenco, katere razvoja najverjetneje ni več mogoče, pa tudi ne smiselno ustaviti. Preteklost nam je pokazala, da ljudje zaradi svojih unikatnih kapacitet nismo enostavno zamenjani, a morda bo napredna UI spremenila tudi to. Trenutno kaže, da imamo resne težave pri nadzorovanju razvoja določenih tehnologij, in nadzorovanju njihovega delovanja. Digitalno izkoriščanje zasebnosti, osebnih podatkov, dostojanstva, in še česa, je privzeta realnost, ki jo marsikdo zavestno ignorira. Medtem se razvijajo vse bolj natančne umetne inteligence, ki vplivajo na izide volitev tudi v najbolj demokratičnih državah, ki nas bodo v kratkem prevažale po cestah in zraku, in ki že obstreljujejo tarče na vojnih območjih – vse to, kot kaže, s precej pomanjkljivim razmislekom o etičnih in legalnih dimenzijah ter posledicah. Teorij ne manjka, strokovnjaki, raziskovalci, filozofi, pravniki, in še marsikdo, že davno svarijo pred vsem, kar se lahko zgodi, če nove tehnologije niso pravilno regulirane, testirane, in na koncu, se razume, odgovorno rabljene. Kot smo omenili v začetku tega prispevka, problem ni toliko tehnološki ali filozofski, kot je sistemski. Do zlorab in negativnih posledic prihaja predvsem za to, ker ljudem v pozicijah z neposrednim dostopom do vseh možnih koristi novih tehnologij, boljša regulacija enostavno ni v političnem ali ekonomskem interesu. Zato je nujno potrebna premestitev teoretičnih etičnih in moralnih premislekov v pravno dimenzijo, kjer

[92] Hunt, E.: Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter, v: The Guardian, 24. 3. 2016.
[93] Povzeto po: Gokmen, A.: Institutional Facts and AMAs in Society, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence, Springer, Leeds 2017, str. 248-251.

bodo lahko nove, še ne popolnoma razumljene tehnologije obravnavane odgovorno in zaščitene pred zlorabami, preden v svet izpustimo nekaj, česar ne bomo zmožni ukrotiti ali ustaviti.

Seznam literature in virov

Monografije

Harcourt, B. E.: Desire and Disobedience in the Digital Age. Harvard University Press, Cambridge 2015.
Tye, M.: Ten Problems of Consciousness: A Representational Theory of a Phenomenal Mind. MIT Press, Cambridge 1995.

Znanstveni članki in poglavja iz knjig

Acemoglu, D. in Restrepo, P.: Artificial Intelligence, Automation, and Work, v: Agarwal, A., Goldfarb, A. in Gans, J.: NBER Working Paper Series (24196). National Bureau of Economic Research, Cambridge 2018.
Akst, D.: Automation Anxiety, v: Wilson Quarterly, 37 (2013) 3, str. 65-77.
Anderson, J. M. in drugi.: Liability Implications of Autonomous Vehicle Technology, v: Anderson, J. M. in drugi: Autonomous Vehicle Technology, RAND Corporation, Santa Monica 2014, str. 111-134.
Anderson, S. L.: Machine Ethics, v: Anderson, J. M in Anderson, S. L.: Machine Ethics. Cambridge University Press, Cambridge 2016, str. 1-19.
Autor, D. H.: Why Are There Still So Many Jobs? The History and Future of Workplace Automation, v: Journal of Economic Perspectives, 29 (2015) 3, str. 3-30.
Balkin, J. M.: How to Regulate (and Not Regulate) Social Media. Uvodni nagovor simpozija Association for Computing Machinery Symposium on Computer Science and Law, New York (2019).
Beckers, S.: AAAI: An Argument Against Artificial Intelligence, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 235-247.
Bellet, T. in ostali: From semi to fully autonomous vehicles: New emerging risks and ethico-legal challenges for human-machine interactions, v: Transportation Research, 63 (2019) F, str. 153-164.
Boddington, P.: Towards a Code of Ethics in Artificial Intelligence, v: Delphi 2 (2019) 2, str. 105-106.
Bodle, R.: The ethics of online anonymity or Zuckerberg vs. 'Moot', v: ACM SIGCAS Computers and Society. 43 (2013) 1, str. 22-35.
Bostrom, N.:Ethical Issues in Advanced Artificial Intelligence, v: Schneider, S.: Science Fiction and Philosophy: From Time Travel to Superintelligence. Blackwell Publishing, Chichester 2009, str. 374-382.
Bostrom, N. in Yudkowsky, E.: The Ethics of Artificial Intelligence, v: Ramsey, W. in Frankish, K.: Cambridge Handbook of Artificial Intelligence. Cambridge University Press, Cambridge 2011, str. 1-20.
Brey, P.: Anticipatory Ethics for Emerging Technologies, v: Nanoethics, 6 (2012) 1, str. 1-13.
Brkan, M.: Freedom of Expression and Artificial Intelligence: on personalisation, disinformation and (lack of) horizontal effect of the Charter, v: MCEL Working Paper Series, Maastricht (2019), str. 1-18.
Brooks, R. in ostali: Is the Brain a Good Model for Artificial Intelligence, v: Nature, 482 (2012), str. 462-463.

Bryson, J. in Winfield, A.: Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems, v: Computer. 50 (2017) 5, str. 116-119.
Burkell, J. in Regan, P. M.: Voter preferences, voter manipulation, voter analytics: policy options for less surveillance and more autonomy, v: Internet Policy Review, 8 (2019) 4, str. 1-24.
Capurro, R.: Digitalization as an ethical challenge, v: AI & Society, 32 (2017), str. 277-283.
Dasoriya, R. in ostali: The Uncertain Future of Artificial Intelligence, v: 8th International Conference on Cloud Computing, Data Science & Engineering, 2018, str. 458-461.
de Sio, F. S.: Killing by Autonomous Vehicles and the Legal Doctrine of Necessity, v: Ethic Theory and Moral Practice, 20 (2017), str. 411-429.
di Norcia, V.: Ethic, Technology Development, and Innovation, v: Business Ethics Quarterly, 4 (1994) 3, str. 235-252.
Gokmen, A.: Institutional Facts and AMAs in Society, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 248-251.
Hacker, P.: Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law, v: Common Market Law Review, 55 (2017), str. 1143-1186.
Hajian, S. in ostali: Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining, v: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 22 (2016), str. 2125-2126.
Kane, T. B.: A Framework for Exploring Intelligent Artifical Personhood, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 255-258.
Laurie, B. in Doctorow, C.: Secure the Internet, v: Nature, 491 (2012), str. 325-326.
Mishra, A.: Moral Status of Digital Agents: Acting Under Uncertainty, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 273-287.
Müller, V. C. in Bostrom, N.: Future Progress in Artificial Intelligence: a Survery of Expert Opinion, v: Müller, V. C.: Fundamental Issues of Artificial Intelligence. Springer, Berlin 2016, str. 553-571.
Nathan, G.: Innovation process and ethics in technology: An approach to ethical (responsible) innovation governance, v: Journal on Chain and Network Science, 15 (2015) 2, str. 119-134.
Peters, M. A., Means, A. J., in Jandrić, P.: Introduction: Technological Unemployment and the Future of Work, v: Peters, M. A., Means, A. J., in Jandrić, P.: Education and Technological Unemployment. Springer, Singapore 2019, str. 1-11.
Prinzing, M.: Friendly Superintelligent AI: All You Need Is Love, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 288-301.
Royakkers, L. in ostali: Societal and ethical issues of digitalization, v: Ethics and Information Technology, 20 (2018), str. 127-142.
Sandvig, C. in ostali: Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, v Data and Discrimination: Converting Critical Concerns into Productive Inquiry. 64th Annual Meeting of the International Communication Association, 2014.
Schwabacher, M. in Goebel, K.: A survey of artificial intelligence for prognostics, v: AAAI Fall Symposium – Technical Report, 2007, str. 107-114.
Statman, D.: Drones and Robots: On the Changing Practice of Warfare, v: Lazar, S. in Frowe, H.: The Oxford Handbook of Ethics and War. Oxford University Press, Oxford 2015.
Swoboda, T.: Autonomous Weapon Systems – An Alleged Responsibility Gap, v: Müller, V. C.: Philosophy and Theory of Artificial Intelligence. Springer, Leeds 2017, str. 302-314.
Thornton, S. in ostali: Incorporating Ethical Considerations Into Automated Vehicle Control, v: IEEE Transactions on Intelligent Transportation Systems, 48 (2017) 6, str. 1429-1439.
Wagner, M. in Koopman, P.: A Philosophy for Developing Trust in Self-Driving Cars, v: Meyer, G. in Beiker, S.: Road Vehicel Automation. Springer, New York 2015, str. 163-171.
Waldrop, M. M.: Autonomous Vehicles: No Drivers Required, v: Nature. 518 (2015) 7537.

Wang, W. in Siau, K.: Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work, and Future of Humanity: A Review and Research Agenda, v: Journal of Database Management, 30 (2019) 1, str. 61-79.

Drugi članki

Heijinen, I.: Fake News Social Media. EuropCom 2017 – Media Literacy Workshop.
Hunt, E.: Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter, v: The Guardian, 24. 3. 2016.
Mozur, P. in Krolik, A.: A Surveillance Net Blankets China's Cities, Giving Police Vast Powers, v: The New York Times, 17. 12. 2019.
Sample, I.: Internet 'is not working for women and girls', says Berners-Lee, v: The Guardian, 12. 3. 2020.

Sodna praksa

C-131/12, Google Spain, ECLI:EU:C:2014:317.
R v Dudley and Stephens (1884) 14 QBD 273 DC.
Re A (conjoined twins) [2001] 2 WLR 480.

Spletni viri

<basilisk.neocities.org> (29. 5. 2020)
<epic.org/privacy/google/glass/#Privacy%20Interests> (29. 6. 2020)